20050930 046

AD 438314

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy)	2. REPORT TYPE	3. DATES COVERED (from. . . to)
August 2005	Interim	January 2003 – December 2004

4. TITLE AND SUBTITLE
Development of Experimental Army Enlisted Personnel Selection and Classification Tests and Job Performance Criteria

5a. CONTRACT OR GRANT NUMBER
DASW01-03-D-0015/DO 0006

5b. PROGRAM ELEMENT NUMBER
622785

6. AUTHOR(S)
Deirdre J. Knapp and Christopher E. Sager(Eds). (Human Resources Research Organization); Trueman R. Tremble (Ed.) (U.S. Army Research Institute)

5c. PROJECT NUMBER
A790

5d. TASK NUMBER
257

5e. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Human Resources Research Organization
66 Canal Center Plaza, Suite 400
Alexandria, VA 22314

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U. S. Army Research Institute for the Behavioral & Social Sciences
2511 Jefferson Davis Highway
Arlington, VA 22202-3926

10. MONITOR ACRONYM
ARI

11. MONITOR REPORT NUMBER
Technical Report 1168

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
Contracting Officer's Representative and Subject Matter POC: Dr. Trueman Tremble

14. ABSTRACT *(Maximum 200 words)*:
U.S. Army leadership recognizes first and foremost the importance of its people – Soldiers – to the effectiveness of transformation to the Future Force. Preparing for this future will affect all aspects of the Soldier management system – selection, job classification, training, and leader development.

This research effort is concerned with Soldier accession and job classification and is titled *New Predictors for Selecting and Assigning Future Force Soldiers (Select21).* The Select21 goal is to ensure the Army acquires Soldiers with the knowledge, skills, and attributes (KSAs) needed for performing the types of tasks envisioned in a transformed Army. The objectives of the project are to (a) identify Future Force job demands and the pre-enlistment KSAs required to meet them, (b) develop measures of job performance and critical KSAs, and (c) validate the experimental predictor (KSA) measures in a concurrent criterion-related validation. This report documents efforts to develop Select21 predictor and criterion measures.

The predictor set includes measures of cognitive ability, temperament, psychomotor skills, values, expectations, and experience. Performance criteria include rating scales to be completed by supervisors and peers, technical knowledge tests, a situational judgment test, and indicators of person-environment fit (e.g., job satisfaction).

15. SUBJECT TERMS

Behavioral and social science	Personnel	Criterion measurement
Selection and classification	Manpower	

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON (Name and Telephone Number)
16. REPORT	17. ABSTRACT	18. THIS PAGE			
Unclassified	Unclassified	Unclassified	Unlimited		Ellen Kinzer Technical Publications Specialist (703) 602-8047

Standard Form 298

Technical Report 1168

Development of Experimental Army Enlisted Personnel Selection and Classification Tests and Job Performance Criteria

Deirdre J. Knapp and
Christopher E. Sager (Eds.)
Human Resources Research Organization

Trueman R. Tremble (Ed.)
U.S. Army Research Institute

Selection and Assignment Research Unit
Michael G. Rumsey, Chief

U.S. Army Research Institute for the Behavioral and Social Sciences
2511 Jefferson Davis Highway, Arlington, Virginia 22202-3926

August 2005

Army Project Number	**Personnel Performance**
622785A790	**and Training Technology**

Approved for public release; distribution is unlimited.

FOREWORD

The U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) conducts research to support Army personnel and training goals. In recognition of the changes emerging with the Army's transformation, ARI developed a research program to identify, describe, and address future personnel requirements. This report describes an aspect of an ongoing ARI project concerned with future enlisted Soldiers.

The objective of this project, known as Select21, is to provide personnel tests for use in selecting and assigning entry-level Soldiers to future jobs. Development of such tests started with a future-oriented job analysis that identified the job performance requirement of future first-term Soldiers and the knowledge, skills, and other personal attributes important for effective performance of the job requirements. The present report describes development of candidate personnel tests based on results of the job analysis. These tests are now being assessed for their validity; that is, the extent to which they predict indicators of effective performance of future jobs. Because future jobs do not yet exist, validity assessment required development of measures of future performance effectiveness. This report also describes the development of the performance criterion measures being used in the validation effort.

Project Select21 is being conducted with support from the Army G-1, Deputy Chief of Staff for Personnel, and from the Army Training and Doctrine Command (TRADOC). ARI has briefed these sponsors, as well as representatives of other offices to include the Army Accessions Command, Human Resources Command, and the Army G-3, Deputy Chief of Staff for Operations. Research sponsors have provided the support and guidance needed for the success of the research.

MICHELLE SAMS
Technical Director

Acknowledgements

There are a large number of individuals and groups not listed as authors who have contributed significantly to the work described in this report. Specifically, we want to acknowledge the contributions of the Select21 Subject Matter Expert Panel and the Scientific Review Panel. The Subject Matter Expert Panel includes the following individuals:

- CSM Cynthia Pritchett
- Mr. James Bierwirth
- LTC Steve Jones
- SGM Rollie Russell
- MAJ Jake Biever
- SFC Ronald Wolfe
- MSG Curtis Rucker
- CW3 Steve Cangiano
- CSM Dan Elder
- Mr. Tom Foster

The Scientific Review Panel includes Wally Borman, Fred Oswald, Bruce Orvis, Ken Pearlman, and Ben Schneider, We are also indebted to the military and civilian personnel who supported our data collection efforts, particularly those Soldiers and NCOs who participated in the instrument development, pilot, and field testing activities.

DEVELOPMENT OF EXPERIMENTAL ARMY ENLISTED PERSONNEL SELECTION AND CLASSIFICATION TESTS AND JOB PERFORMANCE CRITERIA

EXECUTIVE SUMMARY

Research Requirement:

The Select21 project was undertaken to help the U.S. Army ensure that it acquires Soldiers with the knowledge, skills, and attributes (KSAs) needed for performing the types of tasks envisioned in a transformed Army. This transformation will involve development and fielding of Future Combat Systems (FCSs) to achieve full spectrum dominance through a force that is responsive, deployable, agile, versatile, lethal, and fully survivable and sustainable under all anticipated combat conditions (U.S. Army, 2001, 2002). However, Army leadership recognizes first and foremost the importance of its people – Soldiers – for the effectiveness of transformation. In this context, the ultimate objectives of the project are to (a) develop and validate measures of critical KSAs needed for successful execution of Future Force missions and (b) propose use of these measures as a foundation for an entry-level selection and classification system adapted to the demands of the 21st century. In the first stage of the project, we conducted a future-oriented job analysis to support the development and validation effort (Sager, Russell, R.C. Campbell, & Ford, 2005). The present report describes how the job analysis results were used to develop a set of job performance criterion measures and experimental selection and classification predictor measures. This report is primarily targeted toward a technical audience interested in the development process and psychometric qualities of the Select21 research instruments.

Procedure:

The research team developed a wide array of measures in an effort to comprehensively capture pre-enlistment qualifications and assess job/organizational fit. Given the Armed Services Vocational Aptitude Battery (ASVAB) addresses cognitive characteristics quite well, the experimental predictor measures focus primarily on non-cognitive characteristics. The measures include two temperament measures (Rational Biodata Inventory, Work Suitability Inventory), two psychomotor tests, and a "Predictor Situational Judgment Test." We also developed several instruments designed to predict job and organizational fit through the assessment of interests and work-related values. Though it will not be used in the concurrent validation, we developed a prototype measure that could potentially be used for giving Army applicants credit for relevant pre-enlistment education and training.

The research team developed criterion measures designed to, inasmuch as possible, comprehensively forecast performance in future jobs. The measures include rating scales to be completed by supervisors and peers, job knowledge tests, and attitudinal surveys (covering job satisfaction, organizational commitment, etc.). In addition to Army-wide measures, we developed MOS-specific rating scales and job knowledge tests for Soldiers in six target military occupational specialties (MOS) – 11B, 19D, 19K, 31U, 74B, and 96B. Although it will not be possible to use attrition as a criterion for the concurrent validation sample, we will collect separation data from Soldiers who participated in earlier data collections that are described in this report.

The predictor and criterion measures were developed using strategies suitable to their content. For example, the job knowledge test blueprints were based on findings from the future-oriented job analysis and we based test questions on information contained in Soldiers' training manuals. The content of the Work Preferences Inventory, in contrast, was based on a theoretical understanding of career interests (i.e., the Holland model; Holland, 1985). We collected pilot and field test data on the predictor measures from new recruits. We also conducted an additional data collection with new recruits to examine the impact of intentional response distortion (faking) and coaching on several of the predictor measures.

Because of the high degree of deployment activity in the 2003-2004 time period, we had limited access to noncommissioned officers (NCOs) and Soldiers in units to develop and try out the criterion measures. The measures were developed primarily with the assistance of training instructors and were field tested with Soldiers and supervisors in operational units.

Findings:

Over 1,100 new recruits participated in pilot testing of the predictor measures and another 800 participated in the faking research data collection. The field test (conducted August – September 2003) was the first opportunity to administer all the predictors to a sample of new recruits, and it involved almost 700 participants. The predictor measures exhibited good psychometric characteristics and a sensible pattern of score interrcorrelations.

The only opportunity to collect criterion data was in the criterion field test. The goal was to administer the MOS-specific and Army-wide measures to at least 100 Soldiers in each of the target MOS, plus administer just the Army-wide measures to a mixed MOS sample. We collected data from June through October 2004 on a grand total of only 339 Soldiers, with more than 100 cases for just a single MOS (11B). The sample size was sufficient for evaluating the psychometric properties of the Army-wide measures (which were generally quite good) and the data collection procedures (e.g., rater training), but was insufficient for thoroughly evaluating the MOS-specific criterion measures. Additional input from Army subject matter experts was used to prepare the MOS-specific criterion measures for the concurrent validation.

Because we expect circumstances to be similar in 2005, the project team decided to scale back the concurrent validation plan. The most significant change is to collect data on Soldiers in just two (as opposed to six) MOS; this is in addition to an Army-wide (mixed MOS) sample. Therefore, the present report documents field test data analyses pertinent to just the Army-wide, 11B, and 31U (now known as 25U) criterion measures.

Utilization and Dissemination of Findings:

The predictor and criterion instruments described in this report will be administered to Soldiers and their supervisors in a concurrent validation effort planned for 2005. As mentioned, we have reduced the scope of the research to include an Army-wide sample and Soldiers in two MOS. Additional data pertaining to the usefulness of the predictor measures will be available from an "attrition analysis" database. This database includes Soldiers who participated in the predictor pilot, field test, and faking research data collections.

DEVELOPMENT OF EXPERIMENTAL ARMY ENLISTED PERSONNEL SELECTION AND CLASSIFICATION TESTS AND JOB PERFORMANCE CRITERIA

TABLE OF CONTENTS

List of Appendices

List of Tables

List of Figures

DEVELOPMENT OF EXPERIMENTAL ARMY ENLISTED PERSONNEL SELECTION AND CLASSIFICATION TESTS AND JOB PERFORMANCE CRITERIA

CHAPTER 1: INTRODUCTION

Deirdre J. Knapp
HumRRO

Overview of the Select21 Project

The U.S. Army is undertaking fundamental changes to transform into the Future Force. The current project (Select21) concerns future entry-level Soldier selection, with the goal of ensuring the Army selects and classifies Soldiers with the knowledge, skills, and attributes (KSAs) needed for performing successfully in a transformed Army. The ultimate objectives of the project are to (a) develop and validate measures of critical attributes needed for successful execution of Future Force missions and (b) propose use of the measures as a foundation for an entry-level selection and classification system adapted to the demands of the 21st century. The Select21 project focuses on the period of transformation to the Future Force—a transition envisioned to take on the order of 30 years to complete. The time frame of interest extends to approximately 2025.

The major elements of the approach to this project are (a) future-oriented job analysis, (b) development of KSA/predictor measures, (c) development of criterion measures, and (d) concurrent criterion-related validation. The future-oriented job analysis (Sager, Russell, R.C. Campbell, & Ford, 2005) provided the foundation for the development of new tests that could be used for recruit selection or Military Occupational Specialty (MOS) assignment (i.e., predictors) and the development of job performance measures that will serve as criteria for evaluating the predictors. Project researchers will evaluate the potential usefulness of the experimental predictors by comparing Soldiers' scores on the predictor measures to their scores on criterion performance measures in a concurrent criterion-related validation effort.

The Select21 research program is sponsored by the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) with contract support from the Human Resources Research Organization (HumRRO). The ARI/HumRRO project team is supported by three groups. The Army Steering Committee (ASC) comprises Army leaders representing major Army organizations impacted by this research and provides overall direction and support. The Subject Matter Expert Panel (SMEP) includes Army personnel who work with the project team on a more hands-on level (e.g., participating in the job analysis and reviewing measures). The Scientific Review Panel (SRP) is a group of independent experts who periodically review the research plan and findings.

The purpose of this report is to document development of the Select21 predictor and criterion measures. As further background, the remainder of this chapter summarizes the overall Select21 research plan, including the (a) identification of job clusters and job sampling, (b) job analysis findings, (c) criterion measurement plan, (d) predictor measurement plan, and (e) the concurrent validation. The chapter concludes with an overview of the rest of the report.

Job Clusters and Sampling

The Select21 research plan (May, 2002) called for the identification of clusters of future Army jobs. The clusters would provide a basis for determining whether any of the experimental predictor measures had potential for improving classification decisions without relying too heavily on the Army's current job structures (i.e., MOS and associated MOS categorizations such as Career Management Fields [CMF]). The original idea was that the concurrent validation would involve three research samples—an Army-wide sample (with Soldiers from MOS drawn from across each cluster) and samples for at least two job clusters (with Soldiers sampled from MOS within each of those clusters).

As part of the Select21 job analysis effort, 16 future entry-level Army job clusters were identified (Sager et al., 2005). We selected two clusters for closer examination in the validation research: Close Combat and Surveillance, Intelligence, and Communication (SINC). The primary reasons for selecting these two clusters were that they were both considered very important to the Future Force while also being maximally distinct from each other, thus maximizing the opportunity to evaluate the classification potential of the predictor measures.

We identified three MOS to represent each of the two clusters. During the course of subsequent job analysis activity, however, it became clear that the dissimilarities among target jobs within these two clusters were too great to permit job analysis at the cluster level. This was especially true given the need to collect sufficiently detailed job requirement information to support development of job performance measures (e.g., multiple-choice tests of technical knowledge). Therefore, we collected job analysis information for Army-wide requirements (applicable to all MOS) and for six individual MOS representing two job clusters (see Table 1.1).

Table 1.1. Select21 Target Job Clusters and MOS

Close Combat	
11B	Infantryman
19D	Cavalry Scout
19K	M1 Armor Crewman
Surveillance, Intelligence, and Communications (SINC)	
31U	Signal Support Systems Specialist
74B	Information Systems Operator/Analyst
96B	Intelligence Analyst

Note. Two of these MOS are currently undergoing a name change – 31U is becoming 25U and 74B is becoming 25B.

Job Analysis Findings

The Select21 job analysis work characterized future entry-level Army enlisted job requirements in several complementary ways. Job requirements were defined in terms of the following:

- Performance Requirements
  - Performance dimensions (Army-wide)

- Common tasks (Army-wide)
- Job tasks/task categories (for each target MOS)
- Anticipated future conditions (Army-wide and for each target job cluster)

- Pre-enlistment KSAs (Army-wide, prioritized by MOS)

The procedure for conducting the future-oriented job analysis is described in detail in Sager et al. (2005). Following is a summary of the results. Note that, for purposes of this research, an entry-level (or first-term) Soldier is defined as someone with 18-36 months time-in-service.

Performance Requirements

Select21 project staff developed draft performance requirements based on previous ARI-related work (e.g., NCO21 and Project A; J.P. Campbell & Knapp, 2001; Ford, R.C. Campbell, J.P. Campbell, Knapp, & Walker, 2000), Army occupational data, field manuals, and information from the Future Force[1] literature. They presented the draft lists to subject matter experts (SMEs) familiar with the Future Force vision and/or their own MOS. A series of workshops was conducted to capture information about Army-wide, cluster, and MOS requirements.

The 19 Army-wide performance dimensions are listed in Appendix A and the 59 Army-wide common tasks are shown in Appendix B. Note that these tasks are those referred to in the first performance dimension (Performs Common Tasks). Appendix C lists the task categories for the six MOS representing the Close Combat and SINC clusters. Table 1.2 lists the anticipated future conditions for all entry-level Soldiers in the Future Force; the cluster-specific conditions are described in Chapter 3.

Table 1.2. Army-Wide Anticipated Future Conditions

Learning Environment: Greater requirement for continuous learning and the need to independently maintain/increase proficiency on assigned tasks.

Disciplined Initiative: Less reliance on supervisors and/or peers to perform assigned tasks.

Communication Method and Frequency: Greater need to function based on digitized instead of face-to-face communication; greater understanding of the common operational picture and increased situational awareness.

Individual Pace and Intensity: Greater need for mental and physical stamina and greater awareness of one's own mental and physiological status; greater task variety.

Self-Management: Greater emphasis on ensuring that Soldiers balance and manage their personal matters and well-being.

Survivability: Improved protective systems, transportation, communication, and medical care will result in an incremental improvement in personal safety.

[1] In 2003, the Army modified its discussion of the future Army to refer to the "Future Force" rather than the "Objective Force."

Pre-Enlistment KSAs

To identify relevant KSAs, the job analysis team reviewed multiple sources. As described by Sager et al. (2005), these sources included the Basic Combat Training list, Project A KSAs, NCO21 KSAs, Soldier21, as well as several other sources. This activity resulted in a list of 48 KSAs relevant to performance of first-term Soldiers in the Future Force. The list was reviewed by Army SMEs and the Select21 Scientific Review Panel. Appendix D contains the final list of Select21 pre-enlistment KSAs. SMEs provided prioritization ratings of the pre-enlistment ratings overall (i.e., Army-wide) and for each target MOS.

Criterion Measurement Plan

Our goal was to develop criterion measures that, taken together, would provide reasonably comprehensive coverage of the criterion space in terms of content and scores that reflect all performance determinants (i.e., declarative knowledge, procedural knowledge and skills, and motivation) (J.P. Campbell, McCloy, Oppler, & Sager 1993). Implicit in our thinking was also the derivation of a performance model such as that developed for Project A (J.P. Campbell & Knapp, 2001). In that research, first-term Soldier performance was characterized by a model with five factors: Core Technical Proficiency, General Soldiering Proficiency, Effort and Leadership, Maintaining Personal Discipline, and Physical Fitness and Military Bearing. Finally, we wanted to include criteria that address person-environment fit considerations such as job satisfaction and organizational commitment.

The Select21 criterion measures thus include the following:

- Performance Rating Scales
- Job knowledge tests
- Archival/self-report information (e.g., awards, disciplinary actions, attrition)
- Criterion situational judgment test (CSJT)
- Army Life Survey

Figure 1.1 illustrates the content coverage provided by this set of criterion measures. This is not to say, however, that each instrument will provide a score for each performance requirement.

A particularly challenging goal of the Select21 criterion measures is for them to reflect how well Soldiers would perform in the Future Force. Obviously, this is something that must be approximated as closely as possible rather than being a fully achievable goal. The following strategies underlie our efforts to examine future performance:

- Base content of tests on future-oriented job analysis results.
- Provide respondents a basis for making predictions about the future.

These strategies are described further below in the context of the applicable criterion instruments.

Army-Wide Performance Dimensions	Rating Scales[a]	Job Knowledge Tests[b]	CSJT	Archival/ Self-report
Performs Common Tasks	X	X		
Solves Problems/Makes Decisions	X			
Exhibits Safety Consciousness	X	(X)[c]		
Adapts to Changing Situations	X		X	
Communicates in Writing	X			
Communicates Orally	X			
Uses Computers	X			(X)[c]
Manages Information	X			
Exhibits Cultural Tolerance	X			
Exhibits Effort and Initiative on the Job	X			(X)[c]
Follows Instructions and Rules	X	(X)[c]		X
Exhibits Integrity and Discipline on the Job	X			(X)[c]
Demonstrates Physical Fitness	X			X
Demonstrates Military Presence	X			
Relates to and Supports Peers	X		X	
Exhibits a Selfless Service Orientation	X			(X)[c]
Exhibits Self-Management	X		X	
Exhibits Self-Directed Learning	X		X	
Demonstrates Teamwork	X		X	

Note. The Army Life Survey is not listed because it was not designed to cover these performance dimensions.
[a]The MOS-specific rating scales will cover MOS-specific task categories; the Future Expected Scales will cover the anticipated future conditions.
[b]The job knowledge tests cover Army-wide and MOS-specific tasks.
[c]Parentheses indicate indirect assessment of the performance dimension.

Figure 1.1. Select21 criterion measures by performance dimensions matrix.

Performance Ratings

Although ratings tend to exhibit a number of problems when used as a criterion measures, they can tap important dimensions of performance comprehensively and also provide perhaps the best indicator of typical (versus maximal) performance. In Select21, we developed rating scales and data collection procedures intended to maximize the information obtained using this measurement method while minimizing the disadvantages.

As described in detail in Chapter 3, two types of rating scales, designed to be completed by supervisors and peers, have been developed. One set of scales requires raters to consider current observed performance whereas the other set of scales requires raters to estimate how well ratees would perform under different sets of conditions expected to characterize the future Army. The rating scale format, training, and rating procedures are designed to (a) minimize rater errors, (b) focus the raters on the rating scale dimension definitions and anchors, (c) minimize common measurement method bias between the current and future ratings, and (d) facilitate the collection of complete ratings data on all target Soldiers.

Job Knowledge Tests

Job knowledge tests were selected as the primary means by which we would collect data regarding task proficiency. Hands-on tests, which would have provided a more direct measure of task proficiency, were not used because of the resources required to administer them. Although job knowledge tests are lower fidelity assessments, they do offer the advantage of relatively comprehensive task coverage. Moreover, Select21 test developers used a variety of item formats (e.g., multiple-choice, drag and drop, ranking, matching) and graphics to minimize reading requirements and otherwise enhance these computer-administered tests. Project staff drafted seven tests (one Army-wide and one for each target MOS) using test blueprints developed using the Select21 job analysis results and subject matter expert (SME) input. Because these tests cover detailed knowledge of how to perform current job tasks (and comparable information cannot be known for future job tasks), they are not future performance measures, *per se*. The test blueprints are, however, based on findings from the future-oriented job analysis. and there is no reason to believe that the acquisition of declarative knowledge in the future will be predicted by different things than acquisition of such knowledge today.

Criterion Situational Judgment Test (CSJT)

In prior research, several of the Army-wide performance dimensions have been successfully embedded in situational judgment tests (e.g., J.P. Campbell & Knapp, 2001; Knapp, Burnfield et al., 2002). The Select21 Criterion Situational Judgment Test (CSJT) presents problem scenarios common to Soldiers reaching the end of their first terms of enlistment, along with several possible response options. Scores on the test are determined on the basis of judgments provided by senior noncommissioned officers (NCOs). As with the job knowledge tests, the dimensions covered by the CSJT are based on a future-oriented job analysis.

Archival/Self-Report Information

The Personnel File Form, at least variations of it, has been used in several ARI research projects since it was originally developed in Project A (J.P. Campbell & Knapp, 2001). The form draws much of its content from the Army's enlisted personnel "Promotion Point Worksheet." Obtaining the information via self-report is quick, accurate, and efficient (Riegelhaupt, Harris, & Sadacca, 1987) and allows collection of additional information that would not otherwise be readily accessible (e.g., recent disciplinary actions). By its nature, the archival/self-report information reflects performance under current Army conditions.

At the outset of the Select21 research program, the Army was interested in developing experimental predictor measures that would predict attrition as well as performance. Although the Select21 project relies on a concurrent research design that does not allow collection of attrition data from the primary validation sample, considerable data were collected from new recruits in the development and field testing of the predictor measures in 2003-2004. During the timeframe of this project, then, it will be possible to examine the relationship between Select21 predictors and attrition from basic training, advanced training, and (for some research participants) operational units. This work is being conducted somewhat independently from the primary research effort, so it is documented more thoroughly elsewhere (Putka, 2004).

Army Life Survey (ALS)

In an effort to address person-environment fit considerations that broaden the goals of an effective selection and classification system beyond job performance, *per se*, Select21 includes self-report measures of those organizational outcomes associated with attrition. The Army Life Survey (ALS) was developed to measure job satisfaction, organizational commitment, perceived stress, perceived fit, turnover intentions, and perceived importance of core Army values. The Future Army Life Survey (FALS) is a shorter instrument that describes various aspects of the Army of the future and asks Soldiers to indicate how these would affect their feelings toward the Army.

Predictor Measurement Plan

A fundamental goal of the Select21 project is to develop experimental selection and classification measures that will (a) predict performance for entry-level Soldiers in the Future Force and (b) add incremental validity over the current system as embodied by the Armed Services Aptitude Battery (ASVAB). The measures we are developing are designed to cover (inasmuch as possible) the KSAs identified in the Select21 job analysis.

The Select21 predictor measures include the following:

Armed Services Vocational Aptitude Battery (ASVAB)
Temperament measures
- Rational Biodata Inventory (RBI)
- Work Suitability Inventory (WSI)

Psychomotor measures
- Target Shoot
- Target Tracking

Predictor situational judgment test (PSJT)
Record of Pre-Enlistment Training and Experience (REPETE)
P-E fit measures
- Work Values Inventory (WVI)
- Work Preferences Survey (WPS)
- Interest Finder Questionnaire (IFQ)
- Army Beliefs Survey (ABS)
- Pre-Service Expectations Survey (PSES)
- Army Work Knowledge Survey (AWKS)

Figure 1.2 shows the coverage these instruments provide of the Select21 pre-enlistment KSAs. First, note that not all KSAs are covered. In particular, the KSAs related to physical abilities (e.g., static strength, dynamic flexibility) are not addressed. Medical enlistment tests generate scores related to some of these KSAs, but they would not be adequate measures for our needs. Development of alternative measures is outside the scope of ARI's mission. Note also that the P-E fit instruments are not included in Figure 1.2 because they are not designed to cover KSAs, *per se*. Finally, as with the criterion measures, each instrument will not produce scores specific to each pertinent KSA. Rather, the content of the instruments has been designed to reflect the KSAs noted in the figure.

KSA	ASVAB	RBI	WSI	PSJT	Psycho motor	REPETE
Oral Communication Skill						
Oral and Nonverbal Comprehension						
Written Communication Skill						
Reading Skill/Comprehension	✓					
Basic Math Facility	✓					
General Cognitive Aptitude	✓					
Spatial Relations Aptitude	✓					
Vigilance						
Working Memory						
Pattern Recognition						
Selective Attention						
Perceptual Speed and Accuracy						
Team Orientation			✓	✓		
Agreeableness		✓	✓	✓		
Cultural Tolerance		✓	✓			
Social Perceptiveness			✓	✓		
Achievement Motivation		✓	✓	✓		
Self-Reliance		✓	✓			
Affiliation		✓	✓			
Potency		✓	✓			
Dependability		✓	✓	✓		
Locus of Control		✓				
Intellectance		✓	✓			
Emotional Stability		✓	✓			
Static Strength						
Explosive Strength						
Dynamic Strength						
Trunk Strength						
Stamina						
Extent Flexibility						
Dynamic Flexibility						
Gross Body Coordination						
Gross Body Equilibrium						
Visual Ability						
Auditory Ability						
Multilimb Coordination					✓	
Rate Control					✓	
Control Precision					✓	
Manual Dexterity						
Arm-Hand Steadiness						
Wrist, Finger Speed						
Hand-Eye Coordination						
Basic Computer Skill						✓
Basic Electronics Knowledge	✓					✓
Basic Mechanical Knowledge	✓					✓
Self-Management Skill				✓		
Self-Directed Learning and Development Skill				✓		
Sound Judgment				✓		

Note. The P-E fit measures are not included because they are not designed to assess KSAs.

Figure 1.2. Select21 predictor measures by KSA matrix.

Baseline Predictors

The current selection and classification system, which relies largely on the ASVAB, is the baseline against which the Select21 experimental predictors will be compared. The ASVAB contains one experimental test—Assembling Objects (AO)—and the following nine operational tests:

- General Science (GS)
- Arithmetic Reasoning (AR)
- Word Knowledge (WK)
- Paragraph Comprehension (PC)
- Auto Information (AI)
- Shop Information (SI)
- Mathematics Knowledge (MK)
- Mechanical Comprehension (MC)
- Electronics Information (EI)

Applicants must meet a minimum score on the Armed Forces Qualification Test (AFQT) that is a composite of four ASVAB tests (AR, MK, WK, and PC) to enter the Army. For MOS assignment, the applicants' ASVAB scores must meet minimum qualifying scores set for each MOS.

Another baseline predictor is educational status (i.e., high school diploma status), which is used by the Army primarily to predict attrition. ASVAB scores and pre-enlistment educational tier will be retrieved from Soldier records for use in the Select21 research.

Temperament Measures

Prior research tells us that ASVAB is a psychometrically strong measure of cognitive aptitude and an effective predictor of job performance in general and task proficiency in particular. Thus, the experimental predictors developed for Select21 emphasize non-cognitive characteristics that are likely to predict the more motivational aspects of performance and turnover. The first several measures described below illustrate different ways to try to tackle the problem of response distortion (i.e., faking) that has long daunted personnel psychologists. In addition to these instruments, the Select21 research team also seriously considered development of another temperament measure (tentatively called the Fitness for Training Diagnostic) that would address the response distortion problem by virtue of being administered and used post-enlistment. That is, results would be used for post-enlistment identification of new recruits at particularly high risk of attrition or other problems so that positive interventions could be pursued. As we considered this idea more thoroughly, however, it became clear that the development and validation effort would exceed project resources. This is an idea, however, that we believe is worth pursuing as a separate effort.

Rational Biodata Inventory (RBI)

The RBI is an instrument that, in various forms, has been used in prior Army research and operational applications (e.g., for selection into Special Forces) for several years. As its name suggests, the RBI is a self-report measure that uses Likert-style response options. It yields scores on several substantive areas (e.g., Achievement Motivation, Hostility to Authority) and a

response distortion scale. In an operational testing context, the response distortion scale could theoretically be used to identify individuals whose scores should not be used for selection decision-making. Moreover, the instrument development process includes strategies for eliminating items that seem particularly subject to distortion.

Work Suitability Inventory (WSI)

The WSI attempts to address the response distortion problem by (a) using items that reflect work preferences rather than temperament, *per se*, (b) using a forced-choice (i.e., ranking) response format, and (c) allowing construction of alternative composite scores geared to the prediction of different criteria both pre- and post-enlistment (e.g., attrition, performance as a Drill Sergeant, performance as a recruiter). As such, the WSI does not provide scores on individual temperament dimensions that would be useful in a selection context primarily because they are fully ipsative. That is, dimension-level scores are constrained by each other (e.g., if you are high on one dimension you must be lower on another) making it difficult to compare scores across individuals. But the potential advantages of this measurement approach make it a useful addition to the Select21 measurement plan.

Predictor Situational Judgment Test (PSJT)

In addition to being useful for performance measurement, the situational judgment test method has been used even more often as an effective predictor measure (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). We thus developed an experimental predictor using this method. The instrument uses civilian problem scenarios that parallel situations commonly experienced by Soldiers during their first few months in the Army. Researchers are experimenting with several ways to score the PSJT, including one method that would yield temperament-like scores. Particularly if this strategy is successful, the PSJT could be viewed as yet another strategy for assessing temperament that deals with response distortion in yet another way that is distinct from the RBI and WSI.

Psychomotor Tests

Prior research has shown that psychomotor tests can be useful for classifying Army applicants into MOS (J.P. Campbell & Knapp, 2001), but previously the technology for large-scale psychomotor testing was limited. Given advances in this technology, Select21 researchers adapted two psychomotor tests originally developed in prior research. The two tests are Target Shoot and Target Tracking.

Record of Pre-Enlistment Training and Experience (REPETE)

Historically, the Army has taken the burden of training all required entry-level job skills for its enlisted personnel. It stands to reason that recognizing prior related training and/or experience could benefit the Army by reducing training requirements (or at least helping to ensure success in training) and benefit applicants by enhancing their enlistment options (in terms of job choices and/or enlistment bonuses). Such a tool could also be particularly helpful when accessing reserve component Soldiers, and personnel moving from other branches of service, who are more likely to have pertinent job skills prior to entry.

The Select21 project contributed to this idea by developing a self-report experimental predictor measure to determine what types of training and experience entry-level Soldiers bring with them to the Army. To develop this measure, project staff reviewed all the Select21 KSAs and constructed questions that query respondents about related training, certifications, and experience. Particular attention was given to computer-related skills. As discussed further in Chapter 12, the nature of this instrument is such that we do not recommend it be included in the concurrent validation, but the field-tested version helps demonstrate the potential value of this type of measure.

Concurrent Validation

The Select21 research plan calls for the administration of the experimental predictors and the criterion measures to first-term Soldiers in 2005. Other than examination of how early pilot versions of some of the predictors correlate with early term attrition (Putka, 2004), this will be the first look at the combined predictive validity of the ASVAB and the experimental predictor measures using the Select21 criterion measures.

Overview of Report

Chapter 2 completes the background part of this report by providing a general description of the data collections and other activities and instruments that supported development of the Select21 criterion and predictor measures. Part II of the report, which includes Chapters 3 through 7, describes the Select21 criterion measures. Part III includes Chapter 8 through 13 which describe the predictor measures. In Part IV, Chapter 14 describes analyses conducted using scores from the full set of criterion and predictor measures (e.g., to examine correlations among scores within the predictor and criterion sets). Chapter 15 summarizes the Select21 instruments, comments upon issues that remain with their use, and reviews plans for the Select21 concurrent validation effort.

CHAPTER 2: SUPPORTING DATA COLLECTIONS

Deirdre J. Knapp and Christopher E. Sager
HumRRO

Introduction

Development of the Select21 criterion and predictor instruments took place over the course of about 2 years and involved a series of data collections using a variety of Army personnel (e.g., new recruits, senior NCOs) and some civilians (e.g., college students).

To keep pace with the project's aggressive timeline (4 years from job analysis through the concurrent validation), staff began development of data collection instruments before the job analysis work was complete. This was easier for the predictor measures than the criterion measures, which were more dependent on detailed job analysis results for determining applicable content. In addition to the predictor and criterion measures, Background Information Forms were developed to collect descriptive information from research participants; versions of these forms suitable for the concurrent validation will also be developed.

ARI requested the U.S. Army Training and Doctrine Command (TRADOC) and the U.S. Army Forces Command (FORSCOM) to provide training instructors, drill sergeants, students, entry-level Soldiers, and field NCOs to support development of the Select21 predictor and criterion measures. TRADOC was able to provide the requested support, but deployments associated with Operation Iraqi Freedom made it difficult for FORSCOM posts to support the requests. We revised our measurement development plans to reflect changes in troop support availability (primarily by relying more heavily on participation from TRADOC-provided personnel). Considerable energy was devoted to adjusting requests for access to Soldiers at FORSCOM installations to reflect the availability of troops, but the constantly shifting deployment schedules made this quite difficult. Despite our best efforts, we were unable to field test all of the criterion measures (e.g., the job knowledge tests for some MOS) and had relatively limited opportunity to collect performance ratings from supervisors.

Overview of Data Collections

Initial Measure Development

Table 2.1 shows the primary data collection visits made during the course of instrument development. Early visits to Forts Jackson, Leonard Wood, and Lewis were devoted primarily to development of the Predictor Situational Judgment Test (PSJT). The remaining visits in Table 2.1 reflect a series of 2-day workshops with Advanced Individual Training (AIT) and One Station Unit Training (OSUT) instructors and drill sergeants in which the NCOs wrote and revised job knowledge test items, helped develop performance rating scales, and filled out several draft instruments. We also had 3-hour sessions with AIT/OSUT students who filled out draft instruments. The students, instructors, and drill sergeants provided input into development of both the predictor and criterion measures.

Table 2.1. Initial Measure Development Site Visits

Post	Participants	Dates
Fort Jackson	Drill sergeants	5 Nov 02
Fort Leonard Wood	Drill sergeants, AIT Instructors	18 Nov 02
Fort Lewis	AIT Instructors	6 Jan 03
Fort Eustis	AIT Students and Instructors	28 - 31 Jan 03 31 Mar - 4 Apr 03
Fort Benning	AIT Students and Instructors	10 - 13 Feb 03 10 - 12 June 03
Fort Gordon	AIT Students and Instructors	10 - 14 Mar 03 5 - 9 May 03
Fort Knox	AIT Students and Instructors	25 - 28 Mar 03 5 - 9 May 03
Fort Huachuca	AIT Students and Instructors	29 - 30 Mar 03 17 - 18 May 03

Supplemental Data Collections and Expert Reviews

Supplemental activities were required to support development of the situational judgment tests and the job knowledge tests. To support development of the PSJT which has civilian scenarios, college students participated in the development and pre-testing of scenarios. Access to these students was provided by George Mason University in February through June 2003. The judgment-based scoring key for the PSJT was based on data collected from Advanced NCO Course (ANCOC) students (E5 and E6 NCOs) in 2004. The CSJT scoring key was based on data collected from senior (E8 and E9) NCOs at the U.S. Sergeant Majors Academy (USASMA) in June 2004.

Limited access to Soldiers and NCOs in operational units constrained development and pre-testing of the job knowledge tests and performance rating scales. Test questions and rating scales were reviewed by AIT/OSUT instructors and other expert reviewers identified by the Army (e.g., personnel from the Academy of Health Sciences at Fort Sam Houston reviewed the first aid test items). Although it was not possible to pilot test these instruments, they were reviewed with a "field perspective" by a small sample of NCOs provided by Fort Hood in February 2004.

Predictor Pilot Testing and Faking Research

We were fortunate to have enough access to new recruits in Army reception battalions to support considerable predictor pilot testing and a "faking" research effort. The purpose of pilot testing was to administer early versions of the instruments to collect preliminary information about them. In addition to statistical data, we paid attention to administration times and respondent reactions. For example, we revised the wording of items based on questions raised by

the new recruits. As shown in Table 2.2, we collected pilot test data on the predictors at three locations for a total of 1,151 participants. The data collection was limited to 2 hours per Soldier, so Soldiers were divided into groups and administered subsets of the instruments. The goal was to collect data on roughly 200 cases per instrument. Although we were interested in developing predictor measures that could be administered by computer, some of the measures were administered in paper-and-pencil form to reduce the number of computers required. In addition to collecting data on response distortion, some participants in the faking research data collections pilot tested the psychomotor tests that had not been available for administration in the earlier data collections.

Table 2.2. Predictor Pilot Testing and Faking Research

Post	Sample Size	Dates
Predictor Pilot Tests	New Recruits	Sep-Nov 03
Fort Knox	393	
Fort Jackson	465	
Fort Benning	293	
Total	1,151	
Faking Research	New Recruits	Jan-Feb 04
Fort Jackson	551	
Fort Knox	250	
Total	801	

In the faking research data collection (conducted at two reception battalions with a total of 801 participants), we administered predictors potentially subject to response distortion or score inflation due to coaching under various conditions (e.g., respond honestly, present yourself positively but try not to get caught being dishonest) to get an understanding of how they might function in an operational setting. In an operational setting, it can be expected that applicants will be motivated to represent themselves in the most positive light possible in order to be selected for entry and that recruiters may be motivated to help them (e.g., through coaching). The "faking" instructions differed by instrument to reflect concerns specific to each. For example, the PSJT instructions told recruits to make themselves look good to the Army and coached them on strategies for doing that. The research plan included a complex design that varied the instruments each group of examinees took so that data could be collected on all the relevant measures while limiting test administration time to 3 hours per examinee. Despite its complexity, the design reflected a relatively simple strategy of having recruits first take the measure under "normal" for-research-only conditions and then to take the measure with special instructions. See Appendix F for specific "faking" instructions associated with each instrument.

Field Testing

The field tests presented the first opportunity for the full sets of predictor and criterion measures to be administered intact, thus allowing an examination of intercorrelations among measures within each set. Although most of the predictor measures had been extensively pilot tested, this was the first administration of the criterion measures. The purpose of the field test was to finalize instrument content (e.g., many instruments needed to be reduced in length) and

administration procedures in preparation for the concurrent validation. Table 2.3 summarizes the predictor and field test data collections. Predictor data were collected from 690 new recruits at two reception battalions (Forts Jackson and Knox). The measures were administered in a 4-hour period. Criterion data were collected from 339 E3 and E4 Soldiers at five locations (Korea, Fort Lewis, Fort Bragg, Fort Campbell, and Fort Hood). This data collection took 8 hours per Soldier.

Table 2.3. Field Test Data Collections

Post	Participants	Dates (2004)
Predictors	New Recruits	
Fort Jackson	492	13 - 18 Aug
Fort Knox	198	11 - 12 Sep
Total	690	
Criteria	E3-E4 Soldiers	
Korea	111	21 Jun - 9 Jul
Fort Lewis	99	19 - 23 Jul
Fort Bragg	70	2 - 6 Aug
Fort Campbell	2	2 - 6 Aug
Fort Hood	57	12 - 15 Oct
Total	339	

Note. Sample sizes are pre-data cleaning. Criterion field test numbers do not include supervisor raters.

Data Collection Procedures

Data collection protocols were established prior to each data collection activity. Starting with the predictor field tests, detailed Test Administrator (TA) manuals were developed for each type of data collection (predictor pilot test, faking research, predictor field test, criterion field test) and formal training was provided to the HumRRO and ARI staff who served as data collectors.

In addition to criterion measures administered directly to participating Soldiers (e.g., job knowledge tests) and performance ratings from peers, the criterion field test included collection of performance ratings from Soldiers' supervisors. Participating installations were asked to provide two supervisors per Soldier to make these ratings. Supervisors who were unable to attend a face-to-face ratings session were given an envelope with the rating materials and instructions for completing them. As discussed further in Chapter 3, the number of supervisors we were able to get using the face-to-face and "leave behind" approaches was disappointing, with the average number of ratings per Soldier substantially less than one.

ARI purchased 55 IBM Notebook computers for use in the Select21 data collection efforts. It was necessary to minimize the number of computers that needed to be set up at each data collection site because of the limited number of computers and the expense of shipping them. The computer administration requirement was minimized by (a) administering some instruments in a paper-and-pencil format and (b) dividing Soldiers into two groups for testing. One group started with a paper-and-pencil session in one room while the other group started with

a computer session in a second room. Then they would switch for the second half of the data collection period. This strategy was used for both predictor and criterion testing.

Every data collection session began with an introduction to the Select21 project, the participants' role in the project, and administration of a Privacy Act Statement. This statement describes how data will be used and that it will be handled confidentially, thus ensuring the informed consent of Soldiers participating in the research data collection.

Database Management

Several job aids were used to help maximize data quality and minimize data loss. In addition to the detailed administration instructions provided in the TA manuals, data collection logs were used to record on-site difficulties and anomalies (e.g., inattentive respondents). Soldier rosters were used to match Soldier names to project identification codes that were used to help track the various data collection instruments. The computerized measures also included quality control capabilities to help ensure collection of complete and accurate data.

The Select21 databases are maintained by a single database manager who is also responsible for providing database documentation for each. The database for each major data collection described in this chapter (predictor pilot test, predictor faking research, predictor field test, criterion field test) includes item-level as well as composite or constructed variables. Determining how data would be cleaned (e.g., criteria for identifying data from inattentive responders) and development of scoring schemes is the responsibility of the lead analyst working with each instrument. However, the following data cleaning steps were followed for all instruments:

- A Soldier's data for a particular instrument was dropped if the Soldier failed to respond to at least 90% of the items.
- Problem logs were used to identify Soldiers with questionable data that should be dropped.
- Soldiers who completed computerized measures too quickly (relative to most other respondents) were dropped.

Database documentation is available for each of these datasets to facilitate current and future analysis of these data.

Description of the Field Test Data Collection Samples

As further background to the remaining chapters, we close this chapter with a detailed description of the predictor and field test samples. Table 2.4 shows the approximate sample sizes for the predictor field test sample. The numbers are approximate because they do not reflect cases that were subsequently dropped during the data cleaning process. Moreover, analyses reported in subsequent chapters will have varying sample sizes depending upon the instrument being analyzed.

Table 2.4. Predictor Field Test Sample Sizes by Subgroup

Subgroup	Participants
Gender	
Male	498
Female	198
Race	
White	443
Black	114
Other	35
Ethnicity	
White Non-Hispanic	385
Hispanic	86

Note. Sample sizes are pre-data cleaning. Total n = 690.

Table 2.5 shows the approximate sample sizes for the Soldiers who were administered field test versions of the Army-wide criterion measures. Sample sizes for supervisor ratings are described in Chapter 3. Table 2.5 includes Soldiers in the six target MOS who also took MOS-specific criterion measures. The MOS-specific counts are provided in Table 2.6. As is immediately evident, the only MOS for which we met our sample size goals (100 cases with complete data) was 11B (Infantryman). This situation led to a change in the research plan that is described briefly below and discussed further in the final chapter of this report.

Table 2.5. Criterion Field Test Sample Sizes by Subgroup

Subgroup	Participants
Gender	
Male	291
Female	47
Race	
White	215
Black	54
Other	33
Ethnicity	
White Non-Hispanic	202
Hispanic	54
MOS Type	
Army-Wide	111
11B Infantryman	128
19D Cavalry Scout	1
19K M1 Armor Crewman	5
31U Signal Support Systems Specialist	29
74B Information Systems Operator/Analyst	40
96B Intelligence Analyst	25

Note. Total n = 339. Throughout this report, 11C Indirect Fire Infantryman Soldiers are treated as 11B Infantryman Soldiers. Sample sizes are pre-data cleaning. Criterion field test numbers do not include supervisor raters.

Table 2.6. Criterion Field Test Sample Sizes by Subgroup for MOS-Specific Measures

Subgroup	Participants					
	11B	19D	19K	31U	74B	96B
Gender						
Male	128	1	5	27	26	15
Female	0	0	0	2	13	10
Race						
White	96	0	3	14	23	19
Black	10	1	1	8	11	3
Other	6	0	1	4	4	0
Ethnicity						
White Non-Hispanic	89	0	2	13	24	17
Hispanic	24	1	1	5	1	3
Total	128	1	5	29	40	25

Note. 11B = Infantryman. 19D = Cavalry Scout. 19K = M1 Armor Crewman. 31U = Signal Support Systems Specialist. 74B = Information Systems Operator/Analyst. 96B = Intelligence Analyst. Sample sizes are pre-data cleaning. Criterion field test numbers do not include supervisor raters.

Our decision rule was to perform subgroup analyses when subgroups contain at least 20 cases. Relevant subgroups for the predictor measures were race/ethnic group (white, black, non-white hispanic) and gender (male, female). Relevant subgroups for the Army-wide criterion measures included race/ethnic group, gender, and MOS-type. Here MOS-type refers to Army-wide, Close Combat, and SINC. Where supported by sample sizes, the relevant subgroups for Army-wide criterion measures are race/ethnic group, gender, and sample (e.g., Army-wide, 11B, 31U). The MOS-specific criterion measure subgroup analyses are conducted for race/ethnic group and gender. Subgroup means and standard deviations are reported in each instrument chapter. Subgroup effect sizes were calculated by taking the mean of the non-referent group (e.g., females, blacks) minus the mean of the referent group (e.g., males, whites), and dividing the resulting quantity by the standard deviation of the referent group.

A Change of Plan

During 2004, the Army was experiencing a particularly high rate of deployment activity. Despite efforts to adapt our data collection procedures and schedule to these conditions, we were unable to collect sufficient field test data on most of our MOS-specific criterion measures. The outlook for 2005 was not any better and postponing the concurrent validation was not an option. Accordingly, we needed to consider (a) how to prepare the criterion instruments for the concurrent validation given limited field test data for some of them and (b) how to adjust the research plan to maximize the likelihood we will collect sufficient data in the concurrent validation to support our research goals. After considering a variety of options, we decided to scale back plans for the concurrent validation to include an Army-wide sample and just two MOS-specific samples – 11B to represent the Close Combat job cluster and 31U to represent the SINC cluster. Note that, although we had more field test data for the other two SINC MOS (74B and 96B), we expect to have comparatively better luck with 31U in 2005.

We have sufficient field test data to prepare the Army-wide and 11B criterion measures for the concurrent validation. We are currently exploring the possibility of administering just the 31U job knowledge test to another 75 or so Soldiers. If additional data cannot be obtained, we will finalize the 31U job knowledge test and rating scales based on the information in hand (provided through data analysis and as part of another review by Army SMEs planned for January 2005). We do not view this as a problem for the rating scales, for which most revisions are not likely to be data based in any case. It will present a problem for the job knowledge test, where revisions based on statistical item analysis is an important development step. Our back-up plan is to administer a longer test than originally planned so we can drop poorly performing items prior to creating criterion scores.

In another deviation from the original research plan, we will limit the MOS-specific concurrent validation samples to one day of testing rather than the one and a half days. This means we will need to reduce administration time for some of the predictors and criteria and/or drop some measures altogether. Decisions about how to approach this were based in part of the analyses of the field test data and are discussed further in Chapter 15.

CHAPTER 3: PERFORMANCE RATING SCALES

Patricia A. Keenan, Teresa L. Russell, Huy Le, David Katkowski, and Deirdre J. Knapp
HumRRO

Background

The performance rating scales to be completed by Soldiers' supervisors and peers are intended to serve as a primary criterion measure in the Select21 concurrent validation. Scores on the rating scales will help determine the extent to which new predictor tests relate to measures of current job performance and expected performance under future conditions.

We developed two types of rating scales:

- *Current Observed Performance Rating Scales (COPRS):* These scales are for rating the Soldier's current performance. We developed Army-wide and MOS-specific versions of these scales.
- *Future Expected (FX) Performance Rating Scales:* These scales are for rating expected Soldier effectiveness under conditions we anticipate will exist in the future. We developed an Army-wide version of these scales, as well as scales for two target job clusters. Cluster-level scales were possible because, for the most part, the anticipated future conditions were similar enough to apply to all MOS within a cluster.

Current Observed Performance Rating Scales (COPRS)

Overview

Over the course of scale development, the goal was to develop scales that would accurately represent job performance, help minimize common rating errors (and maximize the psychometric quality of the ratings), and be maximally usable by raters. Achieving these goals is a balancing act. Detailed statements might represent job performance more accurately but make the scales cumbersome to read and use. To balance these goals, we developed the COPRS (both Army-wide and MOS-specific) in an iterative process, starting with drafts based on Project A (J. P. Campbell & Knapp, 2001) and NCO21 (Knapp et al., 2002), gathering input first from subject matter experts (SMEs; primarily AIT/OSUT instructors who took part in a series of workshops in the first half of 2003), then from the Select21 Subject Matter Expert Panel (SMEP), and finally from NCOs in the field.

Army-Wide COPRS

We reviewed the Army-wide performance dimensions identified during the job analysis (Sager, Russell, R.C. Campbell, & Ford, 2005) with the goal of combining them into fewer scales. There were 19 performance dimensions (see Appendix A) and a 20th dimension (MOS-specific task performance) to parallel the common task performance dimension. Rating this many dimensions would make the rating task rather onerous, particularly since most raters (especially supervisors) would be asked to rate multiple Soldiers on multiple sets of rating scales.

Therefore, we used an iterative process to organize the dimensions into a smaller set. A draft organization was presented in turn to NCOs at Forts Lewis, Benning, and Eustis. At each meeting, the NCOs' comments and suggestions were incorporated into the next iteration. Both the Scientific Review Panel (SRP) and SMEP reviewed the final plan, which organized the 20 performance dimensions into 12 rating scale dimensions.

Project staff reviewed the anchors from relevant rating scales used in Project A (J. P. Campbell & Knapp, 2001) and NCO21 (Knapp et al., 2002) to develop ideas about what information to include in the draft Army-wide scales. SMEs at several sites used the draft scales to rate two Soldiers that they supervised, after which we asked them for suggestions to improve both the content and format of the scales.

The COPRS contain rating scales for the 12 dimensions listed in Table 3.1. The instrument also includes a single overall performance effectiveness scale. Figure 3.1 provides an example of the format used for all of the COPRS (Army-wide and MOS-specific). Each COPRS has four sections: (a) a title and definition of the target area, (b) behavior examples, (c) summary statements of performance levels (i.e., Below Expectations, Meets Expectations, and Exceeds Expectations), and (d) a 7-point rating scale. Finally, the response form provides a "cannot rate" option for raters to use if they could not rate the ratee on a particular dimension.

Table 3.1. The 12 Army-Wide Current Observed Performance Rating Scales Dimensions

A.	Common Task Performance
B.	MOS-Specific Task Performance
C.	Communication Performance
D.	Information Management Performance
E.	Problem Solving and Decision Making Performance
F.	Adaptation to Changes in Missions/Locations, Assignments, and Situations
G.	Exhibits Level of Effort and Initiative on the Job
H.	Demonstrates Professionalism and Personal Discipline on the Job
I.	Supports Peers
J.	Exhibits Tolerance
K.	Demonstrates Personal and Professional Development
L.	Demonstrates Physical Fitness

MOS-Specific COPRS

Development of the MOS-specific scales was more challenging than developing the Army-wide scales because there was not as much information from past projects to guide development of draft scales. The Project A MOS-specific rating scales included only one of the Select21 target MOS (11B), and NCO21 did not use MOS-specific ratings. So, for most of the MOS-specific scales, development started from scratch—identifying dimensions, developing definitions, and writing behavioral anchors for each scale.

	A. *Common Task Performance*		
Target area →	The extent to which the Soldier performs most Common Tasks (e.g., navigation, first aid, weaponry, maintenance) competently and safely and uses computer applications		
Behavior examples →	– Is not able to perform most Common Tasks – Requires constant supervision – Endangers self and/or others through carelessness – Unable to locate information on the Internet	– Performs most common tasks competently – Requires some supervision under difficult conditions – Typically avoids risks and notices hazards – Can locate most information on the Internet	– Performs almost all common tasks extremely effectively – Requires little or no supervision, even under difficult conditions – Takes steps to protect self and others from hazards – Efficiently locates information on the Internet
	Below Expectations	Meets Expectations	Exceeds Expectations
Rating Scale→	1 2	3 4 5	6 7

Figure 3.1. Example COPRS format.

The job analysis (Sager et al., 2005) identified MOS-specific task categories (shown in Appendix C) that we used as a starting point for the MOS-specific COPRS rating dimensions. Multiple SME panels (AIT instructors for each MOS) reviewed the MOS task categories and recommended reorganization of the categories to make them more suitable for use as rating dimensions. Our goal was to ensure distinct and logical rating dimensions.

Project staff used Project A MOS-specific rating scales to show SMEs what the rating scales would look like and to give them examples of the types of behavior that might be included in the MOS rating scales. SMEs generated a definition for each MOS-specific rating dimension, and then broke the definition down into examples of behavior at the three levels of effectiveness. These behavioral examples provided the draft anchors for the scales, which were reviewed and revised in subsequent SME workshops. Field NCOs provided a final review of the COPRS Army-wide and MOS-specific scales. We developed MOS-specific COPRS for six target MOS (11B, 19D, 19K, 31U, 74B, and 96B). The scales contained from five (74B) to nine (31U) rating dimensions.

Future Expected (FX) Performance Rating Scales

Overview

The COPRS focuses raters on how ratees perform specific aspects of their current jobs, and ratings are based on behaviors observed by the rater. In contrast, the FX scales ask raters to predict how well the ratee might be expected to perform under particular sets of conditions. Thus, the rating task differs with regard to the time reference (current vs. future) and the specificity of the performance being rated (specific aspects of performance aspects vs. overall performance).

We used the anticipated future conditions generated in the job analysis phase of the project to develop drafts of the Army-wide and cluster-specific FX scales. This process was modeled on that used in NCO21 (Knapp et al., 2002). The descriptions elaborated on the information in the job analysis to make the scenarios more specific to conditions individual Soldiers are expected to encounter in the future. SMEP members reviewed the forms and assisted in editing the materials.

The SRP then reviewed the rating forms and made several suggestions designed to help raters distinguish between ratings of current performance and anticipated future performance. This was an important issue because in NCO21 (Knapp et al., 2002), there was concern that ratings of future performance were confounded with current performance. Based on the SRP's suggestions, we revised the format of the scales to reduce the amount of required reading and to further focus on requirements as they pertain to individual Soldiers under each of the anticipated future conditions. In addition, we developed a briefing that described anticipated future conditions. The idea was that the briefing would (a) focus raters on future conditions before they began making future ratings and (b) break the rating response set to help raters differentiate between current and future conditions. After these revisions were made, field NCOs reviewed the rating scales and briefing and made suggestions to adjust their content and length.

Army-Wide FX Scales

The Army-wide FX addressed the following four future conditions identified during the job analysis:

- Learning Environment
- Disciplined Initiative
- Communication Method and Frequency
- Individual Pace and Intensity

Two other future conditions identified during the job analysis—Self-Management and Survivability—were not included in the FX scales. (See Table 2.1 for the descriptions of the future conditions.) Self-Management was excluded because we do not expect the construct to change much in the future—supervisors will expect pretty much the same actions as they do currently—and it is specifically covered in the COPRS. We expected Survivability to be mostly dependent on equipment rather than individual behavior, so it was also excluded from the scales.

The FX rating booklet provides a brief description of future conditions and a rating scale for each condition. As an example, the description and rating scale for Individual Pace and Intensity appear in Figure 3.2. Descriptions for all the future conditions are contained in Appendix E. As shown in Figure 3.2, raters made an overall effectiveness rating for each condition using a 7-point rating scale. Raters used another 7-point scale (shown in Figure 3.3) to rate their confidence in the ratings they provided. This scale was used to get feedback from raters about the perceived utility of their future projections.

Condition A: Individual Pace and Intensity Future conflicts are expected to involve intense and sustained operations that will require physical and mental stamina to conduct high paced operation over long periods. Conditions, such as rules of engagement, hostile forces, threat intent and force mission, could change daily. Soldiers might go from a peacetime CONUS environment to full combat activities in a matter of a few days. Here are some of the expectations of Soldiers envisioned for the future: ▪ Soldiers must be capable of cycling between periods of work and rest instantaneously and at unpredictable intervals. ▪ Soldiers will need to maintain focus and commitment when environments, tasks, responsibilities or personnel change. ▪ Soldiers must recognize and respond to mental cues and images (such as icons and graphics) rather than real life sound or visual images. ▪ Soldiers will be required to process information and data flow without becoming overwhelmed, even when tired or stressed. ▪ Soldiers will face a greater variety of tasks as a result of missions and operational environments.

A. Individual Pace and Intensity *How effectively would you expect the Soldier to meet these future requirements?*						
Not likely to meet the Soldier demands described.		Likely to be generally successful, but will struggle to meet the Soldier demands described.			Likely to successfully meet or exceed the Soldier demands described.	
LOW		MODERATE			HIGH	
1	2	3	4	5	6	7

Figure 3.2. Example Army-wide FX rating scale.

Confidence Rating *How confident are you that your ratings accurately reflect the Soldier's ability to meet these future requirements?*						
Not at all confident that my ratings accurately reflect the Soldier's ability to meet future requirements.		Moderately confident that my ratings accurately reflect the Soldier's ability to meet future requirements.			Absolutely confident that my ratings accurately reflect the Soldier's ability to meet future requirements.	
NOT AT ALL CONFIDENT		FAIRLY CONFIDENT			VERY CONFIDENT	
1	2	3	4	5	6	7

Figure 3.3. FX confidence rating scale.

Cluster-Specific FX Scales

Like their Army-wide counterparts, the cluster-specific scales were based on future conditions identified in the job analysis. The Close Combat FX has scales for three cluster-specific future conditions. Because the MOS in the SINC cluster do not overlap as much as those in Close Combat, we developed two sets of SINC scales—one for 31U and 74B and another for 96B. The 31U/74B FX scales contain three future conditions, and the 96B FX has two. The cluster-specific scales use the same format and rating scales as those in the Army-wide FX booklet, including a confidence rating at the end of the booklet. Condition descriptions for the scales that will be used in the concurrent validation (which will include the 11B and 31U MOS only) are shown in Appendix E.

Rater Training/Process

Figure 3.4 provides an outline of the rater training program used in the field test. The training emphasized the importance of making accurate ratings and thinking about a Soldier's strengths and weaknesses. To this end, the training stressed the importance of accurate performance measures to the overall success of the project. It also stressed the notion that the ratings are for research purposes only, to lessen the tendency of raters to "help" their subordinate or buddy by going easy in the ratings. The training focused on the importance of reading the anchors, thinking about a Soldier's relative strengths and weaknesses, and applying that insight to the ratings. Of course, the training also included admonitions about response tendency and evaluation errors. Such strategies have been only moderately successful for reducing rating error in the past. Therefore, so we tried an exercise intended to get raters to actually read the rating scale dimensions and help them interpret those scales properly.

Specifically, the training program for both the Army-wide and MOS-specific COPRS included a performance dimension sorting task designed to familiarize raters with the dimension definitions and assist them in identifying the relative strengths and weaknesses of each ratee. The original sorting task involved giving raters a set of cards on which the rating scale dimensions and anchors were printed, and asking them to rank order the cards for each Soldier so that the top card was the area in which the Soldier's performance was strongest, and the last card was the Soldier's weakest performance area. However, some raters took this process to the extreme—their ratings and rankings had a very strong linear relationship. For example, they might give a rating of 7 for the strongest dimension, even though that Soldier's strong points might best match the middle anchors of the rating scale. We therefore changed the task so that raters sorted the cards into three categories – "Needs Improvement," "Adequate," and "Strong" – to reflect the performance level of the Soldier they were rating. After sorting the cards, raters to recorded the appropriate category on a separate card and complete the COPRS.

Trainers also addressed this potential problem by explaining that the sorting process should help raters identify a Soldier's relative strengths and weaknesses but they should bear in mind that, for a particular Soldier, the "Strong" category may not necessarily merit a high rating. Raters were asked to keep their card sorts in mind while making ratings, but not to allow the sorts to dictate their ratings.

1. Ratings Overview and Rater Training

- Describes the format of the rating scale, walks through the different parts of the scale, emphasizes the importance of reading the statements and matching the Soldier's performance to the statements to make ratings.
- Describes and depicts common rating errors including halo, leniency, central tendency, recency, and stereotyping.
- Stresses the importance of reading the definitions and taking time to think about them.
- Provides instruction for using the "cannot rate" option.

2. Army-Wide Current Observed Performance Rating Scales (COPRS)

- Trains raters to perform the dimension sorting task. Raters receive a deck of cards—one card for each dimension. Raters sort the cards to reflect the first ratee's job performance. That is, they read the scales on the cards, think about the first Soldier to be rated, and sort the cars into three piles for Strong, Adequate, and Needs Improvement to reflect that Soldier's performance.
- Explains that the sorting task was intended as a guide to help make ratings.
- Instructs raters in the use of the Soldier ID cards and the answer sheet.
- Instructs raters to finish the sorting task for all ratees and then begin making ratings on the Army-wide COPRS.
- Emphasizes the importance of reading the anchors and matching the Soldier's performance to the anchors.
- Explains the Overall Effectiveness rating scale.

3. MOS-Specific Current Observed Performance Rating Scales (COPRS)

- Explains that raters are to use the process described above for making MOS-specific COPRS ratings, including the card sorting process.

4. Army-Wide and MOS-Specific Future Expected (FX) Performance Rating Scales

- Delivers futures briefing consisting of seven color slides and narrative.
- Gives specific instructions for making the FX ratings on the rating form.

Figure 3.4. Outline of field test rater training program.

Field Test Method and Sample

The field test of the rating scales had four primary objectives:

- Assess the usability of the administration procedures and training program.
- Calculate composite performance rating scores.
- Evaluate the psychometric properties of the rating scale scores.
- Investigate relations among scores from the different rating scales.

Because of the small sample sizes, revisions to the rating scale instruments were based on our experience with them and additional SME review, rather than statistical analyses.

Administration

Our goal was for each Soldier to be rated by two supervisors and two to four peers. To facilitate the data collection process, we developed (a) an ACCESS database that served as a tool for making peer assignments, tracking needed supervisor raters, and documenting the rating sessions and (b) a number of forms and processes for collecting, recording, and tracking information during the data collection.

Peers were Soldiers who had worked with the ratee for a month or more. At the beginning of their session, Soldiers completed a "Supervisor and Peer Identification Sheet." On it, the Soldier listed peers who were present in the room and who could rate him/her and peers who were present in the room that he/she could rate. We asked Soldiers to identify four peers in each section of the sheet (i.e., raters and ratees), but this was often not possible.

We made peer rating assignments in two steps to maximize the number of peer raters for each Soldier. Using the ACCESS database, we first entered the names of four peers that the Soldier indicated he/she could rate. Then we entered the names of the peers the Soldier indicated could rate his/her performance. In some instances, the Soldiers indicated they could rate and be rated by the same peers. The ACCESS program automatically paired these cases as raters and ratees. Because we wanted to have as many raters and ratees as possible, the program also "inferred" rating pairs. For example, if Private Smith indicated Private Jones could rate her but Jones had not listed himself as a rater for her, the program identified Jones as a rater for Smith. This helped maximize the number of rater-ratee pairs.

We defined a supervisor as an individual who has supervised the Soldier for at least one month, although a longer time was preferred. So, supervisors might include former supervisors or a more senior Soldier/NCO from the Soldier's unit (including official second-line supervisors). In the field test, we requested two supervisors for each ratee as part of the troop support request to Army installations. As a backup, we also asked Soldiers to identify two Supervisor raters and provide contact information for each.

The goal was to have as many supervisors as possible complete rating packets on-site so that we could provide face-to-face rater training and supervise the process. However, we also delivered self-administered "mail-back" packets to supervisors who could not participate while the data collection team was on-site. The mail-back packages contained a description of the Select21 project, instructions for completing ratings (separate MOS-specific and Army-wide only versions), future Army conditions briefing slides with notes, relevant rating scales and answer sheets, and pre-addressed return envelopes. We did not include the card-sorting exercise(s) in the mail-back packages.

Sample Sizes

As described in Chapter 2, we collected ratings data at multiple locations in Korea and at Forts Lewis, Campbell, Bragg, and Hood. We screened the data to identify forms that may have been completed carelessly or scanned incorrectly. We tallied amounts of missing data on the COPRS and

FX and screened out all forms with 10% or more missing data.[2] Table 3.2 provides the numbers of supervisors for Soldiers in the Army-wide and MOS samples. Table 3.3 provides the numbers of peer raters in the Army-wide and each MOS sample.

Table 3.2. Number of Supervisor Raters for Soldiers by Sample

	Number of Supervisor Raters				
Sample	0	1	2	3	Total w/ratings
Army-wide	30	59	21	1	81
11B	70	29	23	6	58
19D	1	0	0	0	0
19K	3	1	1	0	2
31U	5	3	19	2	24
74B	4	17	13	6	36
96B	4	4	7	10	21
Total	117	113	84	25	222

Table 3.3. Number of Peer Raters for Soldiers by Sample

	Number of Peer Raters									
Sample	0	1	2	3	4	5	6	7	8	Total w/ratings
Army-wide	18	21	19	32	17	1	1	1	1	93
11B	43	8	11	24	41	1	0	0	0	85
19D	1	0	0	0	0	0	0	0	0	0
19K	3	2	0	0	0	0	0	0	0	2
31U	3	9	4	8	4	1	0	0	0	26
74B	9	5	3	16	7	0	0	0	0	31
96B	4	5	5	6	5	0	0	0	0	21
Total	81	50	42	86	74	3	1	1	1	258

As indicated in Tables 3.2 and 3.3, the generally low turnout for the field test data collections resulted in a relatively low number of rater-ratee pair even for the Army-wide sample. Although this makes the results rather tentative, we view the field test analyses as an opportunity to consider the analyses we will use in the concurrent validation (when more data will presumably be available) and to work out some of the conceptual issues related to model development.

Deployment History

During development of the rating scales, we considered adding special scales to measure combat performance. However, SMEs felt that combat experience would be reflected in the observed ratings, so we dropped the idea. As it turns out, more than half of the rater/ratee pairs had been deployed together (55.3% of the peers and 66.2% of the supervisors). Over half of the jointly deployed rater/ratee pairs were Soldiers we tested in CONUS who had rotated home after being deployed in Afghanistan or Iraq. The remainder were rater/ratee pairs we tested in Korea.

[2] "Cannot rate" responses were not included in the count of missing data.

Field Test Results

The primary reason for collecting ratings from both supervisors and peers was the idea that, together, they would provide a more comprehensive perspective on Soldier performance than either would alone. As a practical matter, however, combining the supervisor and peer data was also necessary to increase the overall number of ratee-rater pairs and increase the reliability of the ratings criterion scores. Our intention was also to combine supervisor mail-back ratings with those collected on site, although we conducted analyses to confirm that they were of comparable quality before doing so.

Before we combined these sets of data, we analyzed the data set for each rater type on each instrument separately to determine the characteristics of each on its own. These analyses included (a) calculating inter-rater reliability estimates, (b) determining the extent to which each group used the "cannot rate" option on the COPRS[3], and (c) calculating the correlations between ratings collected within and across rater groups (supervisor and peer). After calculating composite scores based on pooled supervisor and peer data, we examined their psychometric properties and subgroup differences. We conducted similar sets of analyses for the Army-wide COPRS and FX Scales. Because of the small MOS-specific sample sizes, we only report a subset of these analyses for the 11B sample.

Army-Wide COPRS

Supervisor and Peer Ratings Analysis

Inter-rater reliability. The first step in analyzing the supervisor COPRS ratings was to determine whether the self-administered mail-back ratings could reasonably be combined with those collected face-to-face.[4] Across 222 ratees, we collected 377 sets of ratings—268 sets on-site and 109 via mail-backs. We computed inter-rater reliability estimates using intraclass correlations (ICC(C, 1); cf. McGraw & Wong, 1996) for each COPRS dimension. These estimates of reliability assuming a single rater, were calculated both with and without the mail-back ratings, as shown in Table 3.4. Because we needed more than one rater to compute the reliability estimate, only data from Soldiers (ratees) with two or more supervisor raters were included in the analysis. As shown, inter-rater reliability estimates for the ratings including the mail-back data are comparable to those calculated using only on-site ratings. This result is similar to earlier findings in the NCO21 project (Knapp, McCloy, & Heffner, 2004) and confirmed the appropriateness of combining the on-site and mail-back ratings in subsequent analyses.

Tables 3.5 and 3.6 show estimated reliabilities for different numbers of supervisor and peer raters, respectively. Even with increasing numbers of raters, reliability estimates for several of the dimensions are fairly low, with the Exhibits Tolerance dimension being among the most problematic. Not surprisingly, the Supports Peers dimension had the lowest reliability estimate for the supervisor raters but was rated more reliably by peers.

[3] Since the future ratings are speculative in any case, raters are not given a "cannot rate" option.
[4] Only supervisors received mail-back packages, so there is no similar analysis for the peer ratings.

Table 3.4. Army-Wide COPRS: Reliability Estimates for Supervisors With and Without Mail-Back Ratings

AW COPRS Dimension	Inter-Rater Reliability Estimates (ICC(C,1)) Without Mail-Back (n = 45-47)	With Mail-Back (n = 97-105)
A. Common Task Performance	.23	.30
B. MOS-Specific Task Performance	.09	.18
C. Communication Performance	.13	.10
D. Information Management Performance	.33	.28
E. Problem Solving and Decision Making Performance	.17	.20
F. Adaptation to Changes in Missions/Locations, Assignments, and Situations	.19	.14
G. Exhibits Level of Effort and Initiative on the Job	.11	.15
H. Demonstrates Professionalism and Personal Discipline on the Job	.03	.09
I. Supports Peers	.06	.02
J. Exhibits Tolerance	.03	.08
K. Demonstrates Personal and Professional Development	.08	.10
L. Demonstrates Physical Fitness	.42	.35
Overall Performance	.33	.38
Average across all COPRS ratings	.16	.18

Table 3.5. Army-Wide COPRS Reliability Estimates for Supervisor Ratings

AW COPRS Dimension	Number of Raters: 1	2
A. Common Task Performance	.30	.46
B. MOS-Specific Task Performance	.18	.31
C. Communication Performance	.10	.17
D. Information Management Performance	.28	.43
E. Problem Solving and Decision Making Performance	.20	.33
F. Adaptation to Changes in Missions/Locations, Assignments, and Situations	.14	.24
G. Exhibits Level of Effort and Initiative on the Job	.15	.26
H. Demonstrates Professionalism and Personal Discipline on the Job	.09	.17
I. Supports Peers	.02	.04
J. Exhibits Tolerance	.08	.14
K. Demonstrates Personal and Professional Development	.10	.19
L. Demonstrates Physical Fitness	.35	.52
Overall Performance	.38	.55
Average across all COPRS ratings	.18	.29

Table 3.6. Army-Wide COPRS Reliability Estimates for Peer Ratings

	Number of Raters			
AW COPRS Dimension	1	2	3	4
A. Common Task Performance	.20	.33	.43	.50
B. MOS-Specific Task Performance	.30	.46	.56	.63
C. Communication Performance	.17	.29	.37	.44
D. Information Management Performance	.17	.29	.38	.45
E. Problem Solving and Decision Making Performance	.18	.30	.39	.46
F. Adaptation to Changes in Missions/Locations, Assignments, and Situations	.14	.24	.32	.39
G. Exhibits Level of Effort and Initiative on the Job	.23	.38	.48	.55
H. Demonstrates Professionalism and Personal Discipline on the Job	.21	.34	.44	.51
I. Supports Peers	.14	.24	.32	.39
J. Exhibits Tolerance	.03	.07	.09	.12
K. Demonstrates Personal and Professional Development	.20	.33	.43	.50
L. Demonstrates Physical Fitness	.40	.57	.67	.73
Overall Performance	.21	.35	.45	.52
Average across all COPRS ratings	.20	.32	.41	.47

Use of "cannot rate" option. The next step was to determine the extent to which raters used the "cannot rate" option to help determine if any particular COPRS dimension was problematic in terms of "ratability." The option was not used frequently. When it was used, peers used it more somewhat more often (average of 3.9%) than did supervisors (average 2.6%). There was also some tendency for raters (both supervisors and peers) who had been deployed with the Soldier to use the cannot rate option more frequently than those who had not been deployed together.

Correlation between supervisor and peer ratings. Correlations between Soldiers' mean peer and supervisor ratings appear in Table 3.7. These correlations are based on mean scores using data provided by all available supervisor and peer raters. The correlations between corresponding dimensions were low to moderate, with the lowest being Adaptation to Changes. Given that we would expect supervisors and peers to have different opportunities to observe many of these performance areas, there do not appear to be any glaring or troublesome inconsistencies in their ratings.

Development of Composite Scores

Our approach to Army-wide COPRS composite formation was to first identify sound confirmatory factor solutions separately for peer and supervisory data and, in turn, use multi-group confirmatory factor analysis (cf. Maurer, Raja, & Collins, 1998) to determine whether the two have similar factor structures. We then formed composites based on the identified factors.

Method. A group of project researchers identified a number of competing models based on past research (e.g., Project A, NCO21, Can-Do vs. Will-Do, cf. Barrick & Mount, 1995). In total, five models, two of which include several hierarchically nested sub-models, were created. These models were tested against the one-factor model that specifies only one general factor underlying all the performance dimensions. Though this one-factor model has often been found

to be underlying rating data in past research (e.g., Knapp et al., 2004), it is likely that the halo effect inherent in the ratings made it impossible to discover the "true" factor structure of performance ratings. In Select21, we attempted to control for halo effect by specifying two rater effects for all the models. As such, observed ratings were specified as including four components: (a) performance construct(s) (depending on the model tested), (b) rater's effect (halo), (c) dimension-specific measurement error, and (d) random response error. These four components were included in the measurement models to be tested. Analyses were conducted separately for supervisor and peer rating data. For purposes of these analyses, we only used data from Soldiers who had at least two raters. We wanted exactly two raters of each type (supervisor and peer) for the analysis. For those Soldiers with more than two raters, we randomly selected data from two of the raters for analysis.

Table 3.7. Correlations Within and Between Peer and Supervisor COPRS Ratings

Dimension	A	B	C	D	E	F	G	H	I	J	K	L	Overall
A	.19												
B	.53/.48	.31											
C	.50/.37	.37/.40	.13										
D	.45/.43	.47/.54	.41/.50	.23									
E	.50/.42	.53/.58	.49/.55	.51/.55	.22								
F	.43/.41	.50/.47	.32/.35	.42/.52	.45/.50	.06							
G	.50/.48	.54/.48	.38/.41	.41/.49	.46/.55	.45/.56	.33						
H	.39/.38	.46/.39	.29/.34	.37/.40	.46/.43	.37/.46	.60/.65	.31					
I	.25/.37	.30/.42	.24/.27	.30/.44	.29/.44	.27/.47	.39/.56	.37.60	.14				
J	.14/.26	.17/.29	.23/.15	.19/.32	.21/.34	.18/.38	.31/.46	.35/.55	.39/.58	.14			
K	.52/.44	.44/.43	.52/.37	.41/.50	.50/.43	.43/.50	.48/.64	.50/.66	.25/.47	.27/.43	.27		
L	.41/.28	.25/.22	.39/.23	.25/.27	.32/.26	.30/.30	.34.47	.39/.42	.21/.30	.16/.22	.41/.47	.46	
Overall	.57/.54	.64/.56	.49/.46	.54/.57	.63/.60	.55/.60	.62.73	.55/.63	.42/.60	.29/.47	.63/.63	.50/.53	.37

Note. n = 159-258. Correlations between vectors of mean peer and supervisor ratings appear on the diagonal. Dimension correlations appear in the lower triangle for peer and supervisor ratings separately. Peer rating correlations appear first (i.e., peer r/supervisor r).

Results. Results suggested that two competing three-factor models are likely to best represent the ratings for both supervisor and peers. The fit indices for these models are quite good, as shown in Table 3.8.[5] Although both models are a better fit than a one-factor model, it is impossible to determine which of the two three-factor models is better because (a) they are not nested within each other and (b) our sample sizes were rather small. Thus, we decided to retain both models for further cross-validation in the concurrent validation.

Table 3.8. Indices of Fit for Confirmatory Factor Analysis for Models 1 and 2

	CFI	RMSEA	SRMR	$\chi^2_{df=450}$
Model 1	.96	.030	.056	511.00
Model 2	.96	.032	.055	518.21

Note. n_{peer} = 195; $n_{supervisor}$ = 107. CFI = Confirmatory Factor Index, RMSEA = Root Mean Square Error of Approximation, SRMR = Standardized Root Mean Square Residual.

[5] These fit indexes are for the multi-group analyses in which factor structure of supervisor and peer ratings were constrained to be the same.

Table 3.9 presents factor loadings of the supervisor and peer rating dimensions for Model 1. Similar information for Model 2 is shown in Table 3.10. As can be seen from the tables, the two models propose the same first factor (Technical Proficiency and Problem Solving [TPPS]) underlying dimensions A, B, C, D, E, and F. The remaining two factors in the models are different.

Table 3.9. Estimated Standardized Loadings of Army-Wide COPRS Dimensions on the Model 1 Latent Constructs

	Factor: Supervisors			Factor: Peers		
AW COPRS Dimension	TPPS	EI	Teamwork	TPPS	EI	Teamwork
A. Common Task Performance	.33	*.00*[a]	*.00*[a]	.49	*.00*[a]	*.00*[a]
B. MOS-Specific Task Performance	.42	*.00*[a]	*.00*[a]	.37	*.00*[a]	*.00*[a]
C. Communication Performance	.30	*.00*[a]	*.00*[a]	.46	*.00*[a]	*.00*[a]
D. Information Management Performance	.39	*.00*[a]	*.00*[a]	.41	*.00*[a]	*.00*[a]
E. Problem Solving and Decision Making Performance	.26	*.00*[a]	*.00*[a]	.45	*.00*[a]	*.00*[a]
F. Adaptation to Changes in Missions/Locations, Assignments, and Situations	.13	*.00*[a]	*.00*[a]	.29	*.00*[a]	*.00*[a]
G. Exhibits Level of Effort and Initiative on the Job	*.00*[a]	.22	*.00*[a]	*.00*[a]	.31	*.00*[a]
H. Demonstrates Professionalism and Personal Discipline on the Job	*.00*[a]	.57	*.00*[a]	*.00*[a]	.31	*.00*[a]
I. Supports Peers	*.00*[a]	*.00*[a]	.47	*.00*[a]	*.00*[a]	.25
J. Exhibits Tolerance	*.00*[a]	*.00*[a]	.48	*.00*[a]	*.00*[a]	.31
K. Demonstrates Personal and Professional Development	*.00*[a]	.35	*.00*[a]	*.00*[a]	.50	*.00*[a]
L. Demonstrates Physical Fitness	*.00*[a]	.25	*.00*[a]	*.00*[a]	.30	*.00*[a]

Note. TPPS = Technical Proficiency and Problem Solving, EI = Effort and Initiative.
[a] These loadings were constrained to be zero.

Model 1 specifies that the G, H, K, and L dimensions are indicators for Factor 2 (Effort and Initiative [EI]), and dimensions I and J reflect Factor 3 (Teamwork). In Model 2, Factor 2 specifies the G, H, I, and J dimensions (Effort and Teamwork [ET]), and Factor 3 is represented by K and L (Physical Fitness and Self Development [PFSD]). Taken together, it appears that five overlapping composites can be created for all the factors suggested by the two models. Factor intercorrelations for Models 1 and 2 are shown in Table 3.11 and 3.12, respectively.

The multi-group analyses for both models indicated that the magnitudes of the loadings of the dimensions on the latent factors and factor intercorrelations are different for supervisor and peer ratings, with the peer raters showing particularly high Model 2 factor intercorrelations. Despite these differences, the finding of configural equivalence between the supervisor and peer ratings (i.e., they have similar underlying factors) suggests it is reasonable to combine these ratings to form composite scores.

Score calculation. While the factor structure analyses used data from two raters of each type (supervisors and peers), this is not necessarily the ideal composition for scores to be used in

the criterion-related validation analyses. The more ratings included in the score composites, the higher their reliabilities will become. Table 3.13 provides inter-rater reliability estimates (McGraw & Wong, 1996) for different combinations of raters. Although the highest reliability could generally be expected from a combination of two supervisors and two peers, it is unlikely that we will be appreciably more successful in the concurrent validation than in the field test obtaining multiple supervisor raters. Thus, it appears that composites formed by ratings from one supervisor and three peers would have acceptable estimated reliabilities. This combination also reflects the data we expect to collect in the concurrent validation (i.e., one supervisor rater and up to four peer raters per Soldier, as explained in the discussion section at the end of this chapter). Therefore, subsequent analyses are based on composite scores calculated using data from one supervisor and three peer raters. If Soldiers had data from more than one supervisor or three peers, raters were randomly deleted to achieve the desired one and three composition. It should be noted, however, that in the concurrent validation, we anticipate using all available data in the analyses.

Table 3.10. Estimated Standardized Loadings of Army-Wide COPRS Dimensions on the Model 2 Latent Constructs

	Factor: Supervisors			Factor: Peers		
AW COPRS Dimension	TPPS	ET	PFSD	TPPS	ET	PFSD
A. Common Task Performance	.34	*.00*[a]	*.00*[a]	.52	*.00*[a]	*.00*[a]
B. MOS-Specific Task Performance	.44	*.00*[a]	*.00*[a]	.42	*.00*[a]	*.00*[a]
C. Communication Performance	.28	*.00*[a]	*.00*[a]	.44	*.00*[a]	*.00*[a]
D. Information Management Performance	.40	*.00*[a]	*.00*[a]	.43	*.00*[a]	*.00*[a]
E. Problem Solving and Decision Making Performance	.27	*.00*[a]	*.00*[a]	.45	*.00*[a]	*.00*[a]
F. Adaptation to Changes in Missions/Locations, Assignments, and Situations	.16	*.00*[a]	*.00*[a]	.31	*.00*[a]	*.00*[a]
G. Exhibits Level of Effort and Initiative on the Job	*00*[a]	.20	*.00*[a]	*.00*[a]	.42	*.00*[a]
H. Demonstrates Professionalism and Personal Discipline on the Job	*.00*	.47	*.00*[a]	*.00*[a]	.34	*.00*[a]
I. Supports Peers	*.00*[a]	.41	*.00*[a]	*.00*[a]	.06	*.00*[a]
J. Exhibits Tolerance	*.00*[a]	.44	*.00*[a]	*.00*[a]	.12	*.00*[a]
K. Demonstrates Personal and Professional Development	*.00*[a]	*.00*[a]	.52	*.00*[a]	*.00*[a]	.49
L. Demonstrates Physical Fitness	*.00*[a]	*.00*[a]	.31	*.00*[a]	*.00*[a]	.28

Note. TPPS = Technical Proficiency and Problem Solving,, ET = Effort and Teamwork, PFSD = Physical Fitness and Self Development.
[a] These loadings were constrained to be zero.

Table 3.11. Correlations of Model 1 Latent Factors

		TPPS	EI	Teamwork
1	TPPS		.24	.09
2	EI	.87		.67
3	Teamwork	-.04	.20	

Note. TPPS = Technical Proficiency and Problem Solving, EI = Effort and Initiative. Correlations under the diagonal are from peer ratings; those above the diagonal are from supervisor ratings. $n_{peer} = 195$; $n_{supervisor} = 107$.

Table 3.12. Correlations of Model 2 Latent Factors

		TPPS	ET	PFSD
1	TPPS		.21	.26
2	ET	.82		.68
3	PFSD	.78	.66	

Note. TPPS = Technical Proficiency and Problem Solving, ET = Effort and Teamwork, PFSD = Physical Fitness and Self Development. Correlations under the diagonal are from peer ratings; those above the diagonal are from supervisor ratings. $n_{peer} = 195$; $n_{supervisor} = 107$.

Table 3.13. Reliability Estimates of the Army-Wide COPRS Composites Under Different Rater Combination Scenarios

Composite	Composite Inter-Rater Reliability					
	One Supervisor	One Peer	One Supervisor + One Peer	One Supervisor + Two Peers	Two Supervisors + Two Peers	One Supervisor + Three Peers
TPPS	.32	.30	.47	.56	.64	.63
EI	.30	.30	.45	.55	.62	.62
Teamwork	.13	.08	.18	.23	.30	.27
ET	.21	.17	.31	.40	.48	.46
PFSD	.26	.34	.41	.53	.58	.62
Overall Effectiveness	.38	.21	.45	.52	.62	.57

Note. TPPS = Technical Proficiency and Problem Solving, EI = Effort and Initiative, ET = Effort and Teamwork, PFSD = Physical Fitness and Self Development.

Army-Wide COPRS Composite Score Descriptive Statistics

Table 3.14 shows the means and standard deviations for the COPRS composite scores based on data from one supervisor and three peer raters. The table also shows intercorrelations among the scores, all of which are statistically significant. Tables 3.15 and 3.16 indicate no significant score differences related to gender or race/ethnicity.

The comparison of composite scores for the Army-wide and 11B samples (see Table 3.17) indicated differences for EI, Teamwork, ET, and PFSD, with 11B Soldiers being rated lower than Soldiers in the Army-wide sample. Given the low sample sizes, we did not look at any other MOS-specific subgroups.

Table 3.14. Means, Standard Deviations, and Intercorrelations for Army-Wide COPRS Composite Scores

Composite Score	*M*	*SD*	1	2	3	4	5
1 TPPS	4.86	0.72					
2 EI	4.84	0.89	.73				
3 Teamwork	5.08	0.76	.59	.59			
4 ET	4.85	0.82	.68	.87	.86		
5 PFSD	4.97	0.95	.67	.93	.46	.69	
6 Overall Effectiveness	5.11	0.78	.81	.78	.61	.72	.70

Note. TPPS = Technical Proficiency and Problem Solving, EI = Effort and Initiative, ET = Effort and Teamwork, PFSD = Physical Fitness and Self Development. $n = 106$. All corrections are statistically significant, $p < .01$ (one-tailed).

Table 3.15. Army-Wide COPRS Composite Scores by Gender

Composite Score	d_{FM}	Male		Female	
		M	*SD*	*M*	*SD*
TPPS	-0.34	4.91	0.71	4.67	0.75
EI	0.02	4.84	0.90	4.86	0.93
Teamwork	-0.14	5.10	0.71	5.00	0.97
ET	0.10	4.95	0.80	5.03	0.92
PFSD	-0.06	4.87	0.94	4.81	1.05
Overall Effectiveness	-0.14	5.13	0.77	5.02	0.87

Note. TPPS = Technical Proficiency and Problem Solving, EI = Effort and Initiative, ET = Effort and Teamwork, PFSD = Physical Fitness and Self Development. $n_{Male} = 87$, $n_{Female} = 18$. d_{FM} = Effect size for Female-Male mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., Males) are listed second in the effect size subscript. None of the effect sizes are statistically significant, $p < .05$ (two-tailed).

Table 3.16. Army-Wide COPRS Composite Scores by Race/Ethnic Group

Composite Score	d_{BW}	d_{HW}	White		Black		White Non-Hispanic		Hispanic	
			M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
TPPS	-0.21	0.04	4.94	0.67	4.80	0.90	4.93	0.68	4.96	0.56
EI	-0.11	0.15	4.84	0.87	4.74	1.10	4.83	0.88	4.96	0.75
Teamwork	-0.39	-0.08	5.15	0.74	4.86	1.00	5.13	0.72	5.07	0.72
ET	-0.15	0.05	4.99	0.78	4.87	1.13	4.97	0.80	5.01	0.62
PFSD	-0.14	0.26	4.84	0.94	4.71	1.13	4.83	0.94	5.07	0.82
Overall Effectiveness	-0.26	0.22	5.15	0.72	4.96	1.05	5.13	0.74	5.29	0.70

Note. TPPS = Technical Proficiency and Problem Solving, EI = Effort and Initiative, ET = Effort and Teamwork, PFSD = Physical Fitness and Self Development. $n_{White} = 66$. $n_{Black} = 17$. $n_{White\ Non\text{-}Hipanic} = 64$. $n_{Hispanic} = 18$. d_{BW} = Effect size for Black-White mean difference. d_{HW} = Effect size for Hispanic-White Non-Hispanic mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., White) are listed second in the effect size subscript. None of the effect sizes are statistically significant, $p < .05$ (two-tailed).

Table 3.17. Army-Wide COPRS Composite Scores by MOS Type

Composite Score	$d_{AW\text{-}11B}$	AW		11B	
		M	*SD*	*M*	*SD*
TPPS	0.39	5.01	0.49	4.73	0.77
EI	**0.71**	5.07	0.80	4.44	0.93
Teamwork	**0.55**	5.23	0.63	4.81	0.82
ET	**0.74**	5.17	0.68	4.56	0.92
PFSD	**0.53**	5.06	0.89	4.56	0.92
Overall Effectiveness	**0.62**	5.26	0.61	4.78	0.74

Note. TPPS = Technical Proficiency and Problem Solving, EI = Effort and Initiative, ET = Effort and Teamwork, PFSD = Physical Fitness and Self Development. $n_{AW} = 34$. $n_{11B} = 29$.. AW = Army-Wide. 11B = Infantryman.. $d_{AW\text{-}11B}$ = Effect size for AW-11B mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* across all Soldiers. Referent group (e.g., 11B) is listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

Army-Wide FX Ratings

We followed much the same analysis approach for the FX ratings as was used with the COPRS data. We analyzed the supervisor and peer data separately to determine the quality and characteristics of each on its own. Then we computed composite scores and performed subgroup analyses.

Supervisor and Peer Ratings Analyses

Inter-rater reliability. We calculated inter-rater reliability estimates for each future condition for the supervisor sample with and without the mail-back ratings, as shown in Table 3.18. Only Soldiers with two supervisor raters are included in this analysis. Unlike the COPRS data, including mail-back data tended to reduce reliabilities of the ratings across the FX scenarios. However, the sample size would be very small without mail-back ratings, so we included them in subsequent analyses.

Table 3.18. Army-Wide FX Scale Reliability Estimates With and Without Mail-Back Ratings

	Inter-Rater Reliability Estimates (ICC(C,1))	
AW FX Future Condition	Without Mail-Back (*n* = 45-47)	With Mail-Back (*n* = 97-105)
A. Individual Pace and Intensity	.31	.26
B. Learning Environment	.20	.13
C. Disciplined Initiative	.38	.23
D. Communication Method and Frequency	.30	.23

Tables 3.19 and 3.20 show reliability estimates for varying numbers of supervisor and peer raters. These are low to moderate estimates, which are comparable to findings with the Army-wide COPRS.

We computed the correlation between Soldiers' mean peer and supervisor ratings. In these analyses, shown in Table 3.21, all data were used. The pattern of condition intercorrelations was reasonably consistent across rater types, and none of the vector correlations are very low.

Table 3.19. Army-Wide FX Reliability Estimates for Supervisor Ratings

	Number of Raters	
AW FX Future Condition	1	2
A. Individual Pace and Intensity	.26	.41
B. Learning Environment	.13	.24
C. Disciplined Initiative	.23	.38
D. Communication Method and Frequency	.23	.38
Average Across All Future Conditions	.21	.35

Table 3.20. Army-Wide FX Reliability Estimates for Peer Ratings

	Number of Raters			
AW FX Future Condition	1	2	3	4
A. Individual Pace and Intensity	.19	.32	.42	.49
B. Learning Environment	.06	.12	.17	.22
C. Disciplined Initiative	.16	.27	.36	.43
D. Communication Method and Frequency	.08	.16	.22	.27
Average Across All Future Conditions	.12	.22	.29	.35

Table 3.21. Correlations Within and Between Peer and Supervisor FX Ratings

FX Condition	A	B	C	D
A. Individual Pace and Intensity	.36			
B. Learning Environment	.60/.63	.27		
C. Disciplined Initiative	.63/.68	.54/.54	.24	
D. Communication Method and Frequency	.44/.58	.59/.56	.50/.55	.19

Note. n = 155-255. Correlations between vectors of mean peer and supervisor ratings appear on the diagonal. Dimension correlations appear in the lower triangle for peer and supervisor ratings separately. Peer rating correlations appear first (i.e., peer *r*/supervisor *r*).

Development of Composite Score

We used confirmatory factor analysis to examine whether there is one general factor underlying ratings of the four future conditions. For the model tested, we also attempted to control for the rater effects by specifying rater factors apart from the future performance factor as described in the previous section. The model fit well, indicating that it is appropriate to create a single composite score of future performance ratings by averaging scores across future conditions and rater type (supervisors and peers). The indices of fit for the FX model are shown in Table 3.22.[6]

Table 3.22. Indices of Fit for Confirmatory Factor Analysis for Army-Wide FX Model

	CFI	RMSEA	SRMR	$\chi^2_{df=24}$
FX Model	.99	.047	.037	30.97

Note. n_{peer} = 167; $n_{supervisor}$ = 93, CFI = Confirmatory Factor Index, RMSEA = Root Mean Square Error of Approximation, SRMR = Standardized Root Mean Square Residual.

As with the Army-wide COPRS, we estimated the reliabilities for different combinations of numbers of peers and supervisors (see Table 3.23). This confirmed our earlier decision to compute the composite score using one supervisor and three peer ratings to be as comparable as possible to the data we expect to collect in the concurrent validation.

[6] These fit indexes are for the multi-group analyses in which factor structure of supervisor and peer ratings were constrained to be the same.

Table 3.23 Reliability Estimates for FX Scores in the Different Rater Combination Scenarios

	Reliability Estimate
FX Composite Formed by One Supervisor	.23
FX Composite Formed by One Peer	.13
FX Composite Formed by One Supervisor and One Peer	.30
FX Composite Formed by One Supervisor and Two Peers	.36
FX Composite Formed by Two Supervisors and Two Peers	.46
FX Composite Formed by One Supervisor and Three Peers	.41

Descriptive Statistics for FX Scores

The mean Army-wide FX composite score across the entire sample was 4.87 (SD = 0.71). As shown in Table 3.24, the mean score for males was significantly higher than for females. There were no significant differences for race/ethnicity (see Table 3.25) or MOS type (see Table 3.26).

Table 3.24. FX Scores by Gender

	d_{FM}	Male		Female	
		M	SD	M	SD
FX Composite	-0.60	4.95	0.67	4.55	0.77

Note. n_{Male} = 87, n_{Female} = 18. d_{FM} = Effect size for Female-Male mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/SD of referent group. Referent groups (e.g., Males) are listed second in the effect size subscript. The effect size is statistically significant, p < .05 (two-tailed).

Table 3.25. FX Scores by Race/Ethnic Group

	d_{BW}	d_{HW}	White		Black		White Non-Hispanic		Hispanic	
			M	SD	M	SD	M	SD	M	SD
FX Composite	-0.21	-0.19	4.94	0.68	4.80	0.74	4.97	0.68	4.84	0.69

Note. n_{White} = 66. n_{Black} = 17. $n_{White\ Non\text{-}Hipanic}$ = 64. $n_{Hispanic}$ = 18. d_{BW} = Effect size for Black-White mean difference. d_{HW} = Effect size for Hispanic-White Non-Hispanic mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/SD of referent group. Referent groups (e.g., White) are listed second in the effect size subscript. The effect size is not statistically significant effect, p < .05 (two-tailed).

Table 3.26. FX Scores by MOS Type

	$d_{AW\text{-}11B}$	AW		11B	
		M	SD	M	SD
FX Composite	0.00	4.83	0.78	4.83	0.56

Note. n_{AW} = 34. n_{11B} = 29. AW = Army-Wide. 11B = Infantryman.. $d_{AW\text{-}11B}$ = Effect size for AW-11B mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/SD across all Soldiers. Referent groups (11B) are listed second in the effect size subscript. The effect size is not statistically significant, p < .05 (two-tailed).

MOS-Specific Ratings

As previously discussed, only one MOS (11B) provided a large enough sample for analysis (see Table 2.5) and even these results must be viewed tentatively because of the small sample size. However, because we are using the field test as a prototype for the analyses to be performed in the concurrent validation, we chose to include the 11B analyses in this report.

11B COPRS

Inter-rater reliability estimates. There were 364 11B rater/ratee pairs available for analysis. Tables 3.27 and 3.28 present the inter-rater reliability estimates (ICCs), assuming varying numbers of raters, for the supervisor and peer ratings, respectively. The estimates for supervisor raters on the eight dimensions are moderate to zero, generally lower than with the Army-wide scales. The peer ratings seemed to be generally more reliable and closer to the estimates associated with the Army-wide scales.

Table 3.27. 11B COPRS: Reliability Estimates for Supervisor Ratings

	Number of Raters	
11B COPRS Dimension	1	2
1. Perform general communications functions	.13	.24
2. Perform first aid	.00	.00
4. Operate and maintain aiming devices	.12	.21
5. Operate and maintain weapons/antitank/hand grenades	.00	.00
6. Perform general navigation functions	.23	.37
7. Perform tactical operations	.00	.00
8. Operate and maintain night vision devices	.00	.00
Average across all 11B COPRS ratings	.07	.12

Table 3.28. 11B COPRS: Reliability Estimates for Peer Ratings

	Number of Raters			
11B COPRS Dimension	1	2	3	4
1. Perform general communications functions	.33	.50	.60	.67
2. Perform first aid	.22	.36	.46	.53
4. Operate and maintain aiming devices	.11	.20	.28	.34
5. Operate and maintain weapons/antitank/hand grenades	.17	.29	.38	.44
6. Perform general navigation functions	.12	.21	.28	.35
7. Perform tactical operations	.16	.28	.37	.44
8. Operate and maintain night vision devices	.07	.12	.17	.22
Average across all COPRS ratings	.21	.34	.44	.51

Use of "cannot rate" option. Table 3.29 shows the numbers and proportions of the "cannot rate" responses for each 11B COPRS rating dimension. As might be expected given the greater specificity of dimension content, raters used the cannot rate option considerably more on these scales than on the Army-wide COPRS. Over half of the raters chose the "cannot rate" option for the dimension "Operate and maintain the Bradley infantry fighting vehicle." This is unusually high, but not unexpected because this is a specialty within the MOS. We excluded this item in subsequent analyses.

Table 3.29. Number and Percent of 11B COPRS "Cannot Rate" Responses

	Rating Dimensions	"Cannot Rate" Responses	
		Number [a]	Percent [b]
1	Perform general communications functions	72	19.8%
2	Perform first aid	53	14.6%
3	Operate and maintain the Bradley infantry fighting vehicle	195	54.6%
4	Operate and maintain aiming devices	48	13.4%
5	Operate and maintain weapons/antitank/hand grenades	58	16.1%
6	Perform general navigation functions	52	14.3%
7	Perform tactical operations	50	13.7%
8	Operate and maintain night vision devices	47	12.9%

[a] These are the numbers of rater-ratee pairs in which the rater selected the "cannot rate" option. [b] These percentages are calculated based on the total number of 364 pairs.

Supervisor and peer rating intercorrelations. Correlations between supervisor and peer ratings appear in Table 3.30. The correlations between Soldiers' mean peer and supervisor ratings were relatively low. The pattern of dimension intercorrelations, however, was reasonably similar.

Table 3.30. Correlations Within and Between Peer and Supervisor 11B COPRS Ratings

COPRS	1	2	4	5	6	7	8
1. Perform general communications functions	*.22*						
2. Perform first aid	.57/.68	*.25*					
4. Operate and maintain aiming devices	.64/.64	.53/.72	*.23*				
5. Operate and maintain weapons/antitank/hand grenades	.62/.54	.60/.55	.61/.70	*.23*			
6. Perform general navigation functions	.59/.53	.58/.47	.48/.71	.70/.60	*-.07*		
7. Perform tactical operations	.49/.47	.53/.56	.54/.72	.63/.71	.60/.71	*.25*	
8. Operate and maintain night vision devices	.51/.46	.56/.39	.40/.69	.63/.74	.59/.63	.39/.68	*.26*

Note. n = 30-78. Correlations between vectors of mean peer and supervisor ratings appear on the diagonal. Dimension correlations appear in the lower triangle for peers and supervisor rating separately. Peer rating correlations appear first (i.e., peer r/supervisor r).

Composite score formation. Because of the small sample, we did not examine the factor structure of the 11B ratings, as done for the AW ratings. We calculated a single composite scores by combining all available ratings from supervisors and peers (there were insufficient cases to use the one supervisor, three peer combination used for the Army-wide measures) and averaging across all

seven remaining dimensions. The mean score was 4.95 (n = 101, SD = 0.76). The 11B sample consists of all males, predominantly whites, so we did not perform further subgroup analyses.

Close Combat FX Scales

The Close Combat cluster included 11B, 19D, and 19K Soldiers. However, because we only had enough data from 11B to conduct any analyses, the results of those analyses are limited to 11B data. As was the case with the 11B COPRS analyses, we were not able to draw any firm conclusions or conduct more detailed analyses on the data.

Inter-rater reliability. Tables 3.31 and 3.32 show the inter-rater reliability estimates for 11B supervisor ratings and peer ratings, respectively. The data for both sets of raters show very low reliabilities, which, as expected, tend to increase with the number of raters.

Table 3.31. Reliability Estimates for Close Combat FX Supervisor Ratings

	Number of Raters	
CC FX Dimension	1	2
1. More variety in weapons, communication, and vehicles	.07	.13
2. Deployment in different configurations	.00	.00
3. Changes in tasks	.04	.07
Average across all FX ratings	.04	.07

Table 3.32. Reliability Estimates for Close Combat FX Peer Ratings

	Number of Raters			
CC FX Future Condition	1	2	3	4
1. More variety in weapons, communication, and vehicles	.00	.00	.00	.00
2. Deployment in different configurations	.00	.00	.00	.00
3. Changes in tasks	.08	.15	.21	.27
Average Across All Future Conditions	.03	.05	.07	.09

Composite formation. As with the 11B COPRS data, we averaged supervisor and peer ratings across all CC FX dimensions to form an overall composite score. We used all available ratings data. The mean score was 5.01 (n = 105, SD = 0.85). We did not perform subgroup analyses.

Relations Across Rating Scales

Finally, we examined relations among the rating scales to help evaluate whether the current and future ratings are reasonably independent of each other and to get an idea of whether the MOS-specific scales provide sufficiently distinct information from the Army-wide scales to make them useful as classification criteria.

Relations Between COPRS and FX Ratings

As shown in Table 3.33, all correlations between the COPRS and FX composite scores are statistically significant. It is likely that halo is the driving force in these correlations. Further analyses were conducted in the confirmatory factor analysis to help determine whether raters could distinguish between current and future performance.

Table 3.33. Correlations between COPRS and FX Composite Scores

Composite	1	2	3	4	5	6
1 TSSP						
2 EI	.73					
3 Teamwork	.59	.59				
4 ET	.68	.87	.86			
5 PFSD	.67	.93	.46	.69		
6 Overall Effectiveness	.81	.78	.61	.72	.70	
7 FX Composite	.69	.55	.40	.48	.51	.59

Note. TPPS = Technical Proficiency and Problem Solving, EI = Effort and Initiative, ET = Effort and Teamwork, PFSD = Physical Fitness and Self Development. All correlations are statistically significant, $p<.05$.

We conducted a confirmatory factor analysis to examine a model that includes all the factors underlying current and future performance ratings. Specifically, we specified a model with three current performance factors (TPSS, EI, and Teamwork), one future performance factor, and two rating method factors (to control for halo effect).[7] This model fit is acceptable (as shown in Table 3.34), further suggesting the appropriateness of the factor structure of the rating data.

Table 3.34. Indices of Fit for Confirmatory Factor Analysis for Combination Model

	CFI	RMSEA	SRMR	$\chi^2_{df=426}$ [a]
FX Model	.90	.042	.080	520.84

$n = 127$, [a]$p < .05$

More importantly, the estimated correlations between current and future performance factors were low to moderate (from .07 to .42), indicating the discriminant validity of the factors. In other words, it appears that raters could distinguish between ratees' current and future performance.

Relations Between Army-Wide and MOS/Cluster-Specific Ratings

A question of interest was whether inclusion of the 11B scores with the Army-wide composites would provide useful information for classification. Although we will not be able to definitively answer this question until we correlate predictor and criterion measures, we calculated the correlations between the composite rating for 11B COPRS and the composite scores of those Soldiers' on the Army-wide measures (see Table 3.35). All Army-wide

[7] This analysis was only carried out on peer rating data because the supervisor sample size was small (n = 107) compared with the number of the model parameters.

composites were significantly correlated to 11B ratings ($p < .01$). The correlations follow the same pattern as the reliabilities of the Army-wide composites, which may indicate that the 11B ratings do not provide additional information for classification. However, this conclusion is tentative, given the relatively low number of data points for 11B. The correlation between the overall composites for the Army-wide and Close Combat FX scales is .66 ($n = 100, p < .01$).

Table 3.35. Correlations Between Army-Wide Composites and 11B Composite Ratings

Army-Wide Composite	11B
Technical Proficiency and Problem Solving (TPPS)	.68
Effort and Initiative (EI)	.61
Teamwork	.38
Effort and Teamwork (ET)	.55
Physical Fitness and Self Development (PFSD)	.58
Overall Composite	.70

Note. n =100. All correlations statistically significant, $p < 0.01$.

Preparation for the Concurrent Validation

SME Review

Particularly given the limited amount of field test data, it was important to have additional SME review of the rating scales to finalize them for the concurrent validation. We asked 11B and 31U SMEs (AIT/OSUT instructors) to review the Army-wide, 11B, and 31U COPRS with regard to three things: (a) combining scales to reduce the number of ratings, (b) softening the low-end anchors, and (c) ensuring the language used is current. Both groups reviewed the Army-wide COPRS and their MOS-specific COPRS.

With regard to the idea of further reducing the number of rating dimensions on the Army-wide COPRS as a strategy for reducing the time it takes to complete the scales, the 11B SMEs suggested combining dimensions G (Exhibits Effort and Initiative on the Job) and K (Demonstrates Personal and Professional Development). The 31U NCOs did not agree with this suggestion, however, so we did not change the number or organization of the rating dimensions.

We had some concerns that the low-end anchors (i.e., a rating of 1 or 2) might be written too harshly and few raters would use them. Review of scale usage shows that endorsement of the low anchors ranged from 2.1% for Information Management Performance to 13.2% for Demonstrated Physical Fitness. Neither group of NCOs felt the low-end anchors needed to be softened, but suggested adding the degree of supervision required as a way of distinguishing between performance levels. Although the SMEs suggested minor wording changes to some of the rating scales, no changes were needed to make the scales current.

Finalizing Rating Scale Content

In addition to minor wording changes based on the SME review, we substantially reorganized and shortened the instructions provided in the Army-wide COPRS booklet. We also

dropped the confidence ratings from the FX Scales. Both peer and supervisor raters indicated a reasonably high degree of confidence in their ratings (approximately 5.13 on a 7-point scale), so we dropped the rating in the interest of shortening the measure. We also edited one set of FX scales to remove references to 74B, since only 31U Soldiers will be included in the concurrent validation.

Rater Training

During the field tests, we found that re-sequencing the rater training activities slightly made the sessions run more smoothly and made those changes for concurrent validation. The rater training was also revised to correspond to the revised instructions in the Army-wide COPRS booklet.

Troop Support Requests

The low on-site turnout led us to change our plan for collecting ratings in the concurrent validation. We have requested one supervisor for each Soldier, with the expectation we will be able to combine ratings from the single supervisor with data from three to four peers. This will also make the tasking easier for Army installations to accommodate. We still plan to collect distance ratings from supervisors to achieve the numbers of raters we need.

Mail-Back Ratings

We will continue to collect mail-back ratings from supervisors as needed in the concurrent validation to help ensure a supervisor rater for each participating Soldier. We do not plan to extend this peer raters since we were fairly successful collecting peer ratings on-site during the field test and feel that the distance ratings process would not work well with peers.

We considered gathering the distance ratings electronically (via email and/or the Internet), but chose to retain the paper-based system. Although the data collected in this manner is not as "clean" as data collected on-site (e.g., there are errors and missing data), there is little development cost associated with them. We could control the quality of the distance data better if we used automated data collection, but this option is more expensive to develop and introduces a number of psychometric issues (McCoy, Carr, Marks, & Mbarika, 2004). Online ratings also require supervisors to have ready access to the Internet during work hours, which we expect would be an issue for some. We also expect there is a good deal of variability in familiarity and use of the Internet, which would likely to affect whether a rater would actually complete the process (Tomsic, Hendel, & Matross, 2000). After considering the alternatives, we concluded that collecting ratings on-line would create more problems (including the requirement for pilot testing) than it would solve.

The return rate for mail-back packets was much lower than we had expected. In Korea, the POC (an E7) followed up and encouraged supervisors to complete their ratings. The key to success in collecting distance ratings in any form will be to have support from local command and POCs who will follow up and persuade the raters to cooperate. We have found that the most effective plan is to have raters return their completed packets to the POC, who then forwards them for analysis.

Discussion

Scores from the four sets of rating scales examined here (Army-wide COPRS, Army-wide FX Scales, 11B COPRS, and Close Combat FX Scales) showed reasonable distributional properties—that is, they showed acceptable levels of variance. The inter-rater reliability estimates are lower than desired (ranging from .27 to .63 on the Army-wide COPRS composites and .41 for the Army-wide FX composite), but we expect these will improve in the concurrent validation when we will have refined our data collection procedures.

We took particular care to help Soldiers differentiate between current and anticipated future performance. The futures briefing described the anticipated Army-wide changes, and the raters were further instructed to read the FX scales carefully because they contained additional information about the anticipated future. The confirmatory factor analysis of the Army-wide data resulted in a model that includes all the factors underlying current and future performance ratings. This model fit well, and the low to moderate estimated correlations between current and future performance factors indicated that raters could distinguish between ratees' current and future performance. We will retain the briefing for the concurrent validation.

CHAPTER 4: JOB KNOWLEDGE CRITERION TESTS

Maggie Collins, Huy Le, and Lori Schantz[8]
HumRRO

Background

We developed job knowledge tests to serve as a criterion to validate the predictors of performance of first-term Soldiers in the 21st century. Job knowledge tests have an important advantage over other performance measurement methods in that they can efficiently and comprehensively cover job knowledge requirements and they are objectively scored. The obvious difficulty in creating a job knowledge criterion test for future performance requirements is that it is impossible to measure Soldiers' knowledge of performance in areas (a) where Soldiers are not yet required to perform or (b) that do not yet exist. As a result, we have focused on what Soldiers currently know. We hypothesize that Soldiers' ability to acquire declarative (textbook) knowledge will be the same, regardless of what that knowledge is at a given time. Consequently, the job knowledge instruments developed for the Select21 project focus on what Soldiers currently know and perform.

We developed seven job knowledge tests: one Army-wide test applicable across MOS and one test for each of the six target MOS. Some test items in the MOS tests were tracked so that, based on how Soldiers answered the job experience questions, Soldiers in the same MOS might answer different sets of test items.

The development of the job knowledge tests was a three-stage process. The first stage involved identifying the knowledge that should be tested and creating test blueprints that designate the content of each test and the degree to which different content areas are reflected on the tests. The second stage was the development of the test items. The final stage involved field-testing the items. The remainder of this chapter describes these stages in detail.

Test Development

The basis of the test content came from the future-oriented Select21 job analysis (Sager, Russell, R.C. Campbell, & Ford, 2005). Originally we hoped to develop three tests, one Army-wide test to apply to all first-term Soldiers and two cluster-specific tests that would apply to Soldiers in the Close Combat and SINC job clusters. Based on review of the job analysis results, however, it became apparent that there was insufficient overlap in the performance requirements across MOS in a single job cluster. Therefore we created separate tests for the three MOS within each job cluster. As a result, we developed one Army-wide test, and MOS-specific tests for Infantrymen (11B), Cavalry Scouts (19D), and Armor Crewmen (19K) (the Close Combat MOS) and Signal Support Systems Specialists (31U), Information System Operator Specialists (74B), and Intelligence Analysts (96B) (the SINC MOS).

---

[8] In addition to the authors of this chapter, the following individuals contributed to the development of these tests: Karen Moriarty, Carrie Noble, Brian Katz, Roy Campbell, Shonna Waters, Amity Hoenisch, Sonia Kim, and Art Paddock.

Test Blueprints

Identifying Initial Test Specifications

A test blueprint specifies the content areas that are measured by a test as well as the degree to which those content areas are covered on the test. Project staff initially reviewed the Army-wide task categories (shown in Appendix B) and the MOS-specific task categories (shown in Appendix C) to identify areas appropriate for assessment on a written test. The task categories (and constituent tasks) served as the foundation for the test blueprints, with the assumption that Soldiers must have the underlying knowledge to perform the task in order to correctly answer the questions.

Task requirements do not apply uniformly to all Soldiers within the same MOS. Different Soldiers within the same MOS may use different weapons and/or vehicles depending on the unit to which they are assigned. We handled this in part in the job analysis, by specifying tasks at a relatively high level of abstraction (in comparison to the highly detailed tasks traditionally used by the Army). We also did a small amount of tracking—that is, depending on which assignments Soldiers are in, they are given different test items. To facilitate this tracking during testing, Soldiers answer specific items about what weapon, equipment, and/or vehicle they are assigned. Three MOS tests were developed with tracked items: 11B (for Infantry Fighting Vehicles [IFV], for which we developed items for Bradley Vehicles), 19K (for Weapons and Fighting Vehicles), and 31U (for FBCB2 [Future Battle Command Brigade/Below] Equipment).

Revising and Weighting the Blueprint Areas

Advanced Individual Training (AIT)/One Station Unit Training (OSUT) instructors reviewed and revised the draft blueprints in a series of schoolhouse site visits. AIT/OSUT instructors from multiple MOS reviewed the Army-wide blueprint, and instructors in the applicable MOS reviewed the MOS-specific blueprints. Based on their input, we made further revisions to the content of the blueprints. Revisions included combining or subdividing task categories and revising or combining task statements. Some changes were needed to make the future-oriented tasks more suitable for testing in today's Army. Reviewers in the 31U MOS also dropped a security related task category because the content relates to information to which non-military personnel are not allowed access.

Another issue was the overlap of task requirements on the 11B and Army-wide examination. The subject matter experts (SMEs) indicated that 11B knowledge requirements for so-called "common" tasks are more extensive and require higher levels of expertise. This led us to retain some of the same task categories on both blueprints, with the understanding that the11B test questions would be written to require more depth of knowledge. The Army-wide item developers would then focus on the more generic and basic applications of these areas. Also, the tasks within categories common to the 11B and Army-wide blueprints were in some cases different. For example, the 11B blueprint includes tasks for weapons specific to 11B Soldiers.

To determine the weight of each task category on a given test, our SMEs allocated 100 points across all task categories. The final weight for each task category was equal to the average weight across SMEs. The SMEs then rank ordered the tasks within each category by importance.

Project staff took the average ranking to finalize the task rank order. Depending on the weight of the task category and the number of tasks within it, the highest ranked tasks (the top 1 to 7) were retained as part of the test. Based on feedback during the item development phase, we combined overlapping tasks into a single task, within a task category, so that more tasks could be included on the blueprint. For example, three tasks under the 11B category "Operate and Maintain Weapons" related to operating the M240. These were combined into one task on the final 11B test blueprint. The blueprints for the Army-wide, 11B, and 31U tests are shown in Appendix G. The remainder of this chapter focuses on those three tests, since we no longer plan to administer the other four MOS-specific tests in the concurrent validation.

Item Development

Item Goals and Sources

Test items were developed to align with the blueprint specifications. HumRRO staff, independent consultants, and AIT/OSUT instructors wrote items. We used Perception® to author and store items and to create and deliver the tests. Perception® is a computer-based testing and item-banking software product developed by Questionmark Corporation and licensed to HumRRO.

In addition to traditional, multiple-choice test items, item developers used alternative item formats, such as "check all that apply," "rank order," and "matching." Item developers also included photographs, diagrams, and other visual aids as much as possible to illustrate the equipment or action referenced in the question or for the test taker to manipulate (e.g., "drag and drop" type items). The goal was two-fold: to create a test that was more performance oriented than a traditional multiple-choice test and to minimize reading requirements.

We anticipated that the final Army-wide test would have about 70 items and the final MOS tests would include between 40 and 60 items. We expected some items to be dropped from the item bank during the item review process. Therefore, item developers wrote 2-3 times as many items as needed. Item developers wrote the majority of test items using sources from the Army and the Internet. These sources included:

- Soldiers' Manual of Common Tasks
- IET Soldier's Handbook
- Field Manuals and Training Manuals
- Web pages, including:
  - WWW.ADTDL.ARMY.MIL
  - WWW.USAPA.ARMY.MIL
  - WWW.ARMYSTUDYGUIDE.COM

In addition, many graphics were created specifically for this effort.

Project A items were reviewed for possible inclusion in the current project (J.P. Campbell & Knapp, 2001). SME focus groups read through Project A items and indicated whether or not each item was appropriate for all Soldiers, appropriate for a particular MOS, or not appropriate

for any first-term Soldiers. If two-thirds of the SMEs indicated that an item was appropriate, it was added to the database of potential items for the relevant test.

Project staff also developed job experience questionnaires for each test. The questionnaires asked Soldiers to indicate the "last time you performed or were trained on the following task" for each task in the test blueprint. The primary purpose of these questionnaires was to allow us to examine the relationship between proximity of performance and test taking and test scores.

Item Review

All items were subject to several iterations of review. After being reviewed in-house at HumRRO, items were reviewed by AIT/OSUT instructors. Items received a third level of review by additional content experts (usually independent consultants with extensive military experience). Although we were unable to involve NCOs from units in the item development process, we were able to have a small number of NCOs at Fort Hood review the items towards the end of the development process. Army-wide items also went through a final proponency review for technical content. For example, SMEs from the U.S. Army Medical Command (MEDCOM), Army Medical Department and School (AMEDDS), reviewed the chemical, biological, radioactive, and nuclear warfare (CBRN) items, and SMEs from the U.S. Army Infantry Center and School reviewed items measuring knowledge of survival tasks.

Test Form Construction

Project staff constructed field test forms for administration in 2004. We used as many items as possible, given time constraints, for the field test examination. In general, there were 1.5 times the number of items we expected to be on the test form for the concurrent validation.

The job experience questionnaire preceded each job knowledge test. For those MOS tests with tracking items, the tracking items were asked following the job experience questionnaire and asked Soldiers if they used a particular weapon or vehicle or to choose between two weapons or vehicles. If Soldiers used both weapons and/or vehicles, they were asked to choose the one with which they were most familiar. The computer program was set up so that Soldiers were presented with items based on how they answered the tracking questions.

Field Test

Data Cleaning

Preliminary analysis was conducted of test results to identify any potential problems related to miskeyed items and Soldier response patterns. Soldiers who did not respond to at least 90% of the questions were dropped. For the Army-wide test, 11 cases were dropped because of excessive missing data. Six cases were dropped from the 11B test and two cases were dropped from the 31U test. We also examined the amount of time Soldiers took to complete the examination. Although one Soldier took extremely long and one Soldier was extremely quick relative to most Soldiers, there did not appear to be any anomalies in their responses, so they

were included for the analysis. Final sample sizes were 309 for the Army-wide test, 107 for the 11B test, and 25 for the 31U test.

Score Development

Multiple-Choice Item Analysis

As described earlier, the item pool included both traditional multiple-choice items and "non-traditional" items. For the multiple-choice items, we assigned a score of 1 for a correct response and zero for an incorrect response. We used classical item statistics to analyze these questions. These statistics include the percentage of examinees selecting each response option and the point-biserial correlation between the option selected and total score of all the multiple-choice items. Project staff reviewed the item statistics to identify items that needed to be rekeyed, revised, or dropped from the final test forms. In addition, there was some overlap between the Select21 Army-wide test and a test recently administered to E4 Soldiers in another ARI project (PerformM21; R.C. Campbell, Keenan, Moriarty, Knapp, & Heffner, 2004). We incorporated the PerformM21 item analysis results into our review as well.

Non-Traditional Item Analysis

The non-traditional items (e.g., matching) allow more scoring options (e.g., partial credit) that can "extract" more information from the items. We followed an analysis plan that allowed us to score the non-traditional items and then combine those scores with the unit-weighted multiple-choice items so as to neither underweight nor overweight the information derived by these multi-part items.

Matching items. There are two types of matching items (Budescu, 1988): single matching (number of stimuli = number of response options) and multiple matching (number of stimuli < number of response options). Multiple-matching items are scored by counting the number of correctly matched pairs. (A "pair" is a match between a stimulus and a response option.) Therefore, the potential score for a multiple-matching item with k stimuli ranges from 0 to k. With single-matching items, scoring can also be done by counting the number of correctly matched pairs, but the last two pairs are worth only one point because responses to these pairs are totally mutually dependent. Accordingly, the potential score for a single-matching item with k stimuli (and k response options) ranges from 0 to $k-1$.

Drag-and-drop items. Drag-and-drop items consist of four graphics and four labels. The goal is to use the computer mouse to "drag" each label onto its corresponding picture. For scoring purposes, this type of item is treated as a matching item. Thus, the scoring procedure for drag-and-drop items is the same as that of the matching items.

Check-all-that-apply items. An item of this type includes several stimuli associated with a common stem. Examinees are required to check all the stimuli that are correct, given the common stem. Scoring is done by counting the number of correct responses to all the stimuli (i.e., checked if correct *and* not checked if incorrect). Thus, potential scores for check-all-that-apply item ranges from 0 to k (with k being the number of stimuli).

Ranking items. Very little research exists about this type of item in the literature. As a result, we developed the two scoring procedures described below:

1. Matching-like scoring: the options to be ranked were considered item stimuli. These stimuli were matched with their ranked positions (for example, if option *A* was ranked *3rd*, it is "matched' with the *3rd* position). Scoring was based on the number of correct matching pairs. As in the case of single matching mentioned above, scores for a ranking item with *k* stimuli (i.e., there are *k* options to be ranked) ranged from *0* to *k-1*.

2. Pair-wise comparison scoring: Conceivably, in responding to a ranking item, participants implicitly compare all the options to each other and then create a ranking based on results of such pairwise comparisons. Thus, a ranking item can be considered as combining a series of pairwise comparisons of the options. A *k*-option ranking item entails $k(k-1)/2$ pairwise comparisons of the options. Accordingly, scoring for a ranking item can be done by counting the number of correct pairwise comparisons of the ranking solution produced by an examinee. Potentially scores for a ranking item therefore range from *0* to $k(k-1)/2$. An Excel workbook was created to enable this scoring procedure.[9]

Weight the non-traditional items. Adopting the partial-credit scoring procedures for non-traditional items resulted in assigning relatively more weight to these items in the total score (as compared to the traditional multiple-choice items). As a result, we needed to determine a set of weights to optimally combine the multiple-choice and non-traditional items. These optimal weights serve two purposes: (a) ensuring that items are combined most efficiently to minimize the effect of measurement error, and (b) providing a benchmark for the non-traditional items (against the multiple-choice items) that facilitated final selection of items in accordance with the test blueprints.

The procedure to determine weights for the non-traditional items involved the following steps:

1. Use confirmatory factor analysis to examine a model specifying a common factor underlying the selected multiple-choice items and the non-traditional items. In the model, residuals (i.e., measurement errors) of items belonging to the same content domains as specified in the test-blueprint were allowed to be freely estimated (i.e., not constrained to be zero).

2. Examine model fit. Discard non-traditional items with low loadings. Estimate reliabilities of the retained non-traditional items by squaring their loadings on the latent factor (Drewes, 2000).

3. Following the procedure suggested by Wainer and Thissen (2001, pp. 44-47; also see Drewes, 2000), calculate weights for the non-traditional items that maximize the total test score reliability when combined with multiple-choice items.

---

[9] There was insufficient data to try the pairwise comparison scoring procedure, so in subsequent sections we only report results based on the matching-like procedure. However, we hope to revisit the issue of scoring procedures for ranking items in the concurrent validation when there are more data available.

4. Use the weights obtained above to determine the potential contribution of each non-traditional item in total test score (i.e., points for the non-traditional items in relation to that of the multiple-choice items which is fixed at one), thereby determining the appropriate number of items (both multiple-choice and non-traditional items) in each content domain to meet test-blueprint requirements.

Calculate Composite Score

A composite score for each test was calculated that summed the selected multiple-choice and appropriately weighted non-traditional items. This score is reported as a percentage correct score to facilitate its interpretability.

Descriptive Statistics and Test Intercorrelations

Table 4.1 describes the properties of the composite scores, including the number of items and points each is worth, the mean and standard deviation, and the estimated reliability. We estimated the reliability of the scores using the formula for weighted composite scores (Feldt & Brennan, 1989; Wainer & Thissen, 2001).

Table 4.1. Properties of the Job Knowledge Tests

Job Knowledge Scores	Scale Properties					
	Number of Items	Number of Nontraditional Items	Maximum Point	Mean	*SD*	Reliability
Army-Wide	64	2	65	60.57	11.17	.77
11B	42 (46)[a]	6	57 (61)[a]	66.96	14.00	.83
31U	45 (48)[b]	4	49 (53)[b]	52.71	14.67	NA[c]

Note. Scores are reported as percents of the maximum points [a] For the 11B test, Soldiers in units that use IFV had to respond to 4 additional items. So the maximum point for those recruits is 61. [b] For the 31U test, some Soldiers had to respond to 3 additional items (out of those, there is one ranking item worth 2 points). So the maximum point for these Soldiers is 53. [c] Reliability for the 31U test could not be estimated because of the small sample size.

For Soldiers who took the Army-wide test and either the 11B or 31U test, we computed correlations between the two sets of scores. The Army-wide and 11B tests were significantly correlated .71 (n = 81), and the Army-wide and 31U tests were significantly correlated .59 (n = 20).

Subgroup Comparisons

We examined subgroup differences when subgroups contained at least 20 cases. Using this rule, comparisons were made among Army-Wide examinees based on gender, race, ethnicity (Hispanic versus White, Non-Hispanic), and MOS (Army-wide, 11B, 31U). A comparison was made between 11B test takers based on ethnicity (Hispanic versus White, Non-Hispanic), but not on gender or race. We made no subgroup comparisons among 31U respondents because the numbers were not sufficient. Results of subgroup comparisons are shown in Tables 4.2 through 4.4.

Table 4.2. Job Knowledge Scores by Gender

Job Knowledge Scores	d_{FM}	Male		Female	
		M	*SD*	*M*	*SD*
Army-Wide	**-0.81**	61.78	10.37	53.40	13.03

Note. [a] $n_{Male} = 218$, $n_{Female} = 37$. d_{FM} = Effect size for Female-Male mean difference. Effect size is calculated as (mean of non-referent group – mean of referent group)/*SD* of the total group. Referent group (i.e., Males) is listed second in the effect size subscript. The effect size is significant, $p < .05$ (two-tailed).

Table 4.3. Job Knowledge Scores by Race/Ethnic Group

Job Knowledge Scores			White		Black		White Non-Hispanic		Hispanic	
	d_{BW}	d_{HW}	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Army-Wide [a]	**-0.91**	**-0.40**	63.25	11.02	53.17	9.25	63.26	10.91	58.91	10.00
11B [b]		-0.19					68.72	13.30	66.23	15.95

Note. [a] $n_{White} = 167$. $n_{Black} = 36$. $n_{White\ Non\text{-}Hispanic} = 157$. $n_{Hispanic} = 41$. [b] $n_{White\ Non\text{-}Hispanic} = 66$. $n_{Hispanic} = 16$. d_{BW} = Effect size for Black-White mean difference. d_{HW} = Effect size for Hispanic-White Non-Hispanic mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., White) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

Table 4.4. Job Knowledge Scores by MOS

Job Knowledge Scores	$d_{AW\text{-}11B}$	$d_{AW\text{-}31U}$	$d_{31U\text{-}11B}$	AW		11B		31U	
				M	*SD*	*M*	*SD*	*M*	*SD*
Army-Wide	**-0.39**	0.37	**-0.76**	60.00	11.57	64.38	9.97	55.92	7.97

Note. $n_{AW} = 92$. $n_{11B} = 92$. $n_{31U} = 21$. AW = Army-Wide. 11B = Infantryman. 31U = Signal Support Systems Specialist. $d_{AW\text{-}11B}$ = Effect size for AW-11B mean difference. $d_{AW\text{-}31U}$ = Effect size for AW-31U mean difference. $d_{31U\text{-}11B}$ = Effect size for 31U-11B mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* across all Soldiers. Referent groups (e.g., 11B) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

As the tables illustrate, we found statistically significant differences among subgroups on the Army-wide test. Soldiers in 11B did significantly better on the Army-wide test than other AW test takers and, specifically, 31U. This is somewhat likely to be due to the similarity in knowledge requirements between the 11B and Army-wide task categories. This similarity may in part also explain the significant difference between male and female performance on the Army-wide test, as females do not have combat roles in the U.S. military. The differences in the race and ethnic group comparisons reflect typical effect size differences observed in cognitive based tests (e.g., Hunter & Hunter, 1984; Jensen, 1980).

Impact of Job Experience

We performed correlation analyses to determine if there is a relationship between how Soldiers answered the job experience questions and how they performed on portions of the test measuring related content areas. These analyses examined two questions: (a) Is there a relationship between whether or not Soldiers have any job experience or training in a task and how they performed on that content area of the test and (b) does the recency of experience or training have an effect on how Soldiers perform in that content area?

Results showed that, in general, correlations between job experience and job knowledge in the same content areas are positive, (ranging from -.08 to .26 with a mean of .10). This finding was expected, and it provides evidence of the construct validity of the job knowledge scales. None of the correlations is high enough to suggest potential problems (e.g., criterion contamination) for using these scales as the criteria in the concurrent validation.

For the Army-wide task, "Process Casualties" which included subtask "Recover and Bury Remains," 80% of the Soldiers indicated that they had never performed or received training on this task. Therefore we dropped that category for the concurrent validation and redistributed the Process Casualties weights proportionately across the remaining blueprint task categories.

The majority of respondents indicated that they had training on or experience in performing the tasks on the 11B job knowledge test, with the exception of the Bradley Fighting Vehicle tasks, which were tracked. There were three questions on the field test where Soldiers' responses to experience were significantly related to performance on the test item. However, each item reflected a different task category and the performance on the remaining items in those categories were not impacted by experience. Therefore all sections of the blueprint were included in the analysis.

Final SME Review

The final scored field test forms include the most psychometrically sound items that correspond to the test blueprint specifications. Items with poor statistics that could be easily corrected (e.g., overlapping response options) were included in the field test results but have been revised for the concurrent validation. AIT instructors at Fort Benning reviewed the 11B test and a subset of the Army-wide items. AIT instructors at Fort Gordon reviewed the 31U test and the remaining Army-wide items. In addition to reviewing items for currency and any performance problems, these SMEs reviewed the blueprint for currency.

Subject matter experts made recommendations for revisions or exclusion of items based on their experience with the content. Minor revisions and recommendations were made regarding the 31U items. In addition to similar minor revisions, 11B AIT instructors recommended that all the "Visual Signaling" items be replaced. The field test items reflected flag signals, which apply more to vehicle drivers than most other Soldiers. Based on this recommendation, HumRRO staff and military consultants developed a new set of items reflecting hand signals. Only those new items with appropriate item statistics will be scored in the concurrent validation test form.

Discussion

In the concurrent validation, we plan to administer the Army-wide test to all participating Soldiers and the 11B or 31U tests (as applicable) to Soldiers in the two target MOS samples. The Army-wide test has 60 questions and the MOS tests have 50 questions each, plus 3 to 4 tracked questions. The inclusion of tracked sections of the MOS examinations (i.e., FBCB2 items in the 31U exam and IFV items in the 11B exam) will depend on where the concurrent validation administrations occur. For example, if the tests are administered in locations where all Infantry Soldiers use Bradley Fighting Vehicles, the IFV items will be included for everyone.

As some items were edited between the field test and the concurrent validation, the psychometric properties may change, presumably for the better. For example, where we identified a distractor (i.e., incorrect response option) that overlapped with the correct response, we scored both as correct on the field test. With edits, however, there should be only one correct answer on the concurrent validation version of those items. We expect this to result in improved psychometric characteristics for the tests compared to the field test results.

CHAPTER 5: CRITERION SITUATIONAL JUDGMENT TEST (CSJT)

Gordon W. Waugh and Teresa L. Russell
HumRRO

Background

A situational judgment test (SJT) measuring aspects of contextual performance and leadership served as a useful criterion measure during the Army's Project A (J. P. Campbell & Knapp, 2001). With that end in mind, we developed the Criterion Situational Judgment Test (CSJT) to measure the job performance of enlisted Soldiers with 18 to 36 months of experience in areas relevant to the Future Force.

Overview of CSJT Development

Our approach to developing the CSJT was a fairly well established one, similar to approaches used by others (e.g., Motowidlo, Dunnette, & Carter, 1990). It had four major steps, each of which involved collecting data from NCOs: (a) scenario generation, (b) response option generation, (c) item review, and (d) scoring key development.

Scenario Generation

We gathered CSJT scenarios at Forts Lewis, Eustis, Benning, and Gordon. During each workshop, we spent about 15 minutes instructing NCOs on how to write SJT scenarios related to specific Select21 performance dimensions. While participants were writing, we circulated through the room to read scenarios and query the writers about their scenarios.

Because the CSJT items ask respondents what *should* be done in a situation, the CSJT taps knowledge and judgment rather than motivation. Therefore, the CSJT measures aspects of *can-do* performance rather than *will-do* performance. Prior to the workshops, we had selected six performance dimensions from the Select 21 job analysis (See Appendix A) that we thought could be assessed using an SJT format. Two performance dimensions, pertaining to teamwork and support for peers, were yielding more potentially useful scenarios than the others. Even so, there were some scenarios that we could make into items for most of the other selected performance dimensions. For the *exhibiting effort and initiative* scenarios, the only appropriate course of action was too obvious. Therefore, we dropped that performance dimension from the test plan for the CSJT. The final five performance dimensions included on the CSJT are listed in Figure 5.1.

The scenarios were typed into a relational database. We recorded several attributes of each scenario, such as its target dimension, relevance to the future military, relevance to the civilian sector, status in the data collection efforts, and so on. In turn, staff assessed the potential usefulness of each scenario obtained in the workshops. When evaluating each scenario, we considered the following characteristics of a well-designed scenario:

- There are several possible response options (i.e., actions) for the scenario.
- There are several possible response options that some people will choose as best.
- The potential response options differ in effectiveness.

- The scenario is likely to be relevant in the future.
- The scenario is relevant to E3 and E4 Soldiers with 18 to 36 months of experience. (A staff member with considerable knowledge in this area made this judgment.)
- The respondent has all the information needed to answer the question.
- The wording is clear and succinct.

If a scenario did not possess these characteristics, we tried to improve it. If we were unsuccessful, we dropped it. At the end of the evaluation process, we had retained 89 CSJT scenarios.

---

1. *Adapts to Changing Situations.* Is able to maintain commitment when environments, tasks, responsibilities, or personnel change. Does not allow stress in high-pressure situations to interfere with job performance. Easily commits to learning new things when the technology, mission, or situation requires it.
2. *Relates to and Supports Peers.* Treats peers in a courteous, respectful, and tactful manner. Shows concern for others by providing help and assistance. Backs up and fills in for others when needed.
3. *Exhibits Self-Management.* Effectively manages own responsibilities (e.g., work assignments, personal finances, family, and personal well being), and appears on duty prepared for work. Sets goals, makes plans, and critically evaluates own performance. Works effectively without direct supervision, but seeks help when appropriate.
4. *Exhibits Self-Directed Learning.* Takes responsibility for mastering skills and learning to apply those skills in the job. As necessary, effectively invests time in learning and practice. Mastery of skills includes those (a) acquired during basic and advanced individual training, and (b) additional skills required by the soldier's initial assignment.
5. *Demonstrates Teamwork.* Understands own and team tasks in relation to the mission or assignment. Coordinates with and helps members maintain focus on the team's goals.

Figure 5.1. Definitions of CSJT performance dimensions.

Response Option Generation

Once the scenarios were written, we asked about 50 E5–E7 NCOs at Forts Benning, Gordon, Knox, and Eustis to describe actions that an E3 or E4 Soldier might take in a given scenario. The vast majority of these NCOs were drill sergeants and Advanced Individual Training (AIT) instructors. For these workshops, the participants completed workbooks containing selected scenarios. The workbooks instructed participants to imagine that they are the person in the situation and to think of at least three response options. To help participants get started writing response options, we gave them the following tips for thinking of actions:

- Think about the action you think would be the best action for the main character to take in that situation.
- Think about the action you think people would take in the situation, even if it is not correct.
- Think about different ways to handle the situation effectively.
- Think about actions you have seen people take in similar situations.

Senior staff members reviewed, condensed, and edited response options with the goals of (a) eliminating or merging redundant statements, (b) clarifying statements, and (c) retaining 6–9 response options for each scenario.

Forty-two E5–E7 NCOs at Forts Eustis, Knox, Gordon, and Benning evaluated the response options by rating the effectiveness of each. Almost all were AIT instructors. Many items were revised and added before Fort Benning, where 18 NCOs (E5–E7 AIT instructors) participated. Using their data, we identified response options with the following desirable characteristics:

- Low standard deviation of effectiveness ratings by NCOs indicating that NCOs agreed with each other on the effectiveness of the response option.
- For each scenario, a set of response options representing a range of effectiveness.

After response options were written, edited, and reviewed, 76 CSJT scenarios had sufficient response options to serve as items and were retained.

Item Review

Ten NCOs participated in a CSJT review workshop in Alexandria, VA. To counteract potential fatigue effects, we assembled two booklets (Form A and Form B). In Form B, the items were in reverse order. Half of the reviewers completed Form A, the other half completed Form B. For each item, reviewers were asked to (a) indicate whether the situation is appropriate for a first-term Soldier (i.e., "an E3–E4 Soldier who has been in the Army for 18–36 months") and (b) comment on any other concerns they might have about the situation. Next, reviewers were asked whether each response option was realistic, appropriate for first-term Soldiers, and factually correct. One senior project staff member with extensive military experience also reviewed the items. After examining the judgments and comments, we dropped 10 items, leaving 66 items for the field test version of the CSJT.

Scoring Key Development

Twenty-six NCOs who were attending USASMA (United States Army Sergeants' Major Academy) participated in a workshop to develop the final scoring key. Each NCO rated the effectiveness of the options in all 66 items retained for the field test. After examining notes from the data collection and individuals' data, we dropped one NCO from the analyses. One additional NCO was dropped because his judgments correlated little with the other NCOs. The interrater reliability among the final NCOs across all options was .97. For the options in the two 33-item forms used in the field test, the interrater reliability was .97 on both forms.

Field Test

Our primary goal was to develop, by the end of the field test, a version of the CSJT with reasonably good psychometric properties that could be administered in one hour or less during the concurrent validation.

Sample and Materials

As described in Chapter 2, we collected CSJT data at several locations in Korea and at Forts Lewis, Campbell, Bragg, and Hood. We screened the data to identify forms that may have been completed carelessly or scanned incorrectly. Next, we tallied amounts of missing data on the CSJT and screened out all forms with 10% or more missing data. One additional Soldier, with 9.5% of responses missing, was also dropped because the Soldier with the next-most missing responses had far fewer missing responses. In all, we retained 321 of the original 330 Soldiers for the analyses.

The CSJT-FT (field test version) had two forms (A and B) with 33 unique items on each form. The items on Form A were intended to cover the performance dimensions of Adaptability and Self-Management. The items on Form B covered the other dimensions: Exhibiting Effort, Relating to Peers, Self-Directed Learning, and Teamwork. Each Soldier completed one of the two forms within the 90 minutes allowed for CSJT testing. Each test item had a stem—a description of a military scenario or situation. The items contained 5–10 response options (i.e., actions that could be taken in the situation), with an average of seven response options. The CSJT-FT asked participants to answer each item by rating the effectiveness of each response option (rather than picking the best and worst options). This response format allows the most flexibility in deciding how to score the items. Soldiers rated the effectiveness of each response option (i.e., action) using the 7-point rating scale shown in Figure 5.2.

Ineffective action. The action is likely to lead to a bad outcome.			Moderately effective action. The action is likely to lead to a passable or mixed outcome.		Very effective action. The action is likely to lead to a good outcome.	
—— Low ——			**——— Moderate ———**		**—— High ——**	
1	2	3	4	5	6	7

Figure 5.2. CSJT response option rating scale.

We computed the judgment score for each response option using Equation 1 below.

$$\textit{Judgment Score}_{Option\ x} = 6 - \left|\, \textit{SoldiersRating}_{Option\ x} - \textit{keyedEffectiveness}_{Option\ x} \,\right| \qquad (1)$$

We subtracted the difference between the rating and keyed effectiveness values from 6 to reflect the scores (i.e., so that higher values would represent better scores). The judgment score for an entire test form was the mean of the option scores.

The *keyedEffectiveness* value for each option was computed from the USASMA subject matter experts' (SMEs') ratings. A typical approach is to use the SME mean rating as the scoring key value. Our research on the Predictor Situational Judgment Test (PSJT), however, found problems with this method (see Chapter 10). The bulk of the keyed values tend to be near the middle of the rating scale because they are means. Therefore, a Soldier coached to rate every option a 4 can achieve a good score by doing so. Because the keyed value is the mean rating among the SMEs, any disagreement on an option whose effectiveness is near an extreme (1 or 7) will move the key value towards the middle of the scale. For example, a keyed value can equal 1.0 only when all SMEs rate the option a 1.

To counteract this effect, the SME means were adjusted by stretching the range of values and rounding to the nearest integer. If the mean SME rating was exactly 4, then no stretching was done. Mean ratings below 4 were lowered slightly, mean ratings above 4 were raised slightly. The amount of change depended on the rating's distance from 4. The farther the rating's distance from 4, the more the rating was changed. Equations 2 and 3 show how the SME mean was adjusted.

$$\textit{if mean} < 4.0\text{: } \textit{AdjustedKey} = \textit{Mean} - .5 * (4 - \textit{Mean}) \tag{2}$$

$$\textit{if mean} > 4.0\text{: } \textit{AdjustedKey} = \textit{Mean} + .5 * (\textit{Mean} - 4) \tag{3}$$

Then the value was rounded to the nearest integer. If an adjusted key value was below 1, it was changed to 1; if an adjusted key value was above 7, it was changed to 7. The resulting scoring key was no longer biased towards the middle of the scale. A Soldier who rates every option a 4 will, thus, get a very low score.

Using the final scoring key, a total score of 6.0 is perfect, and a score of .98 is the lowest possible score. On average, a person responding randomly would achieve a score of 3.6, based on simulated random data.

Judgment Scores

The first step was to examine the psychometric properties of the judgment scores from the two forms. Table 5.1 provides the descriptive statistics and reliability estimates for judgment scores for the two forms.

Table 5.1. Descriptive Statistics and Reliability Estimates by CSJT Form

	n	*M*	*SD*	$k_{response\ options}$	alpha
Form A	156	4.69	0.36	234	.94
Form B	165	4.46	0.39	227	.96

Select Items and Options of the Concurrent Validation Version of the CSJT

The goal was to develop a concurrent validation version of the CSJT (CSJT-CV)—one test form with 27 items. Each item would have four response options, not seven. Our strategy was to identify psychometrically sound response options and then, in turn, move to the item level to identify items for retention.

Option-Level Statistics

For each response option, we computed the means and standard deviations of the option scores and the option total score correlation. Average statistics across response options on the two forms appear in Table 5.2. We removed all response options having near zero option-judgment correlations.

Table 5.2. Option-Level CSJT-FT Statistics

Statistic	Form A	Form B
Number of Items	33	33
Number of Response Options	234	227
Average Option-Judgment Correlation	.25	.28

Item-Level Statistics

For each of the 33 items on each form, we computed the (a) number of retained options for each item, and the (b) average option-total score correlation for the item. Items with fewer than four retained options were eliminated. Then, we selected the final 27 items for the test form by attempting to balance the content dimensions for the scenarios and maximize the average option-total score correlation for the item.

Psychometric Properties of CSJT Scores

Having determined which CSJT-FT options and items should be retained in CSJT-CV, we created two mini-forms—*a* and *b*. Mini-form *a* contained all of the options/items (13 items, 52 options) retained from FT Form A, and mini-form *b* contained all of the options/items (14 items, 56 options) retained from FT Form B. Our intent is that the combined mini-forms will be one test form in the concurrent validation (i.e., CSJT-CV). We analyzed them separately because recruits were nested within forms in the field test.

Descriptive Statistics

We computed means, standard deviations, and reliability estimates for scores on each mini-form. We also used the Spearman-Brown estimate the reliability of scores on CSJT-CV-length versions of mini-forms *a* and *b*. Those data appear in Table 5.3. The two mini-forms correlated highly with the original forms. The correlation was .94 and for both Form A and Form B. The interrater reliability of the scoring key was .98 for both mini-form *a* and mini-form *b*.

Table 5.3. Means, Standard Deviations, and Reliability Estimates of CSJT scores

Score	k	M	SD	r_{kk}	$r_{27\ item}$
Mini-Form *a*	52	4.43	0.53	.88	.94
Mini-Form *b*	56	4.50	0.59	.91	.95

Note. The Spearman-Brown prophecy formula was used to estimate the reliability of a 27-item test with 108 response options. The sample sizes were 156 and 165 for mini-forms A and B, respectively.

The sample sizes were too low to factor analyze the option scores. Therefore, the item scores (13 in mini-form *a*, 14 in mini-form *b*) were factor analyzed, although the sample sizes were still quite small (156 for mini-form *a*, and 165 for mini-form *b*). We ran a parallel analysis (cf. Humphreys & Montanelli, 1975) to determine the number of factors to extract for each mini-form. Parallel analysis computes eigenvalues for random data and compares it to the eigenvalues for the real data. The two scree plots are then plotted. The point where the two scree plots cross determines the number of factors to extract. This procedure was done for 1,000 random datasets.

Mini-form *a* had two factors, and mini-form *b* had four factors. The factors accounted for 40% and 52% of the total variance in mini-forms *a* and *b*, respectively. In a single-factor solution, the lone factor accounted for 36% and 38% of the total variance in mini-forms *a* and *b*, respectively. The factor loadings were not consistent with the performance dimensions. That is, except for some of the Teamwork items, items within a performance dimension did not load on the same factor. This result was expected because SJT items tend to be multidimensional (Motowidlo et al., 1990). Therefore, scale scores were not computed for the CSJT. Subsequent analyses used only the overall CSJT scores.

Tables 5.4, 5.5, and 5.6 report gender, racial/ethnic, and MOS subgroup differences in CSJT scores. MOS differences are fairly large, with 31U getting the highest scores and 11B getting the lowest scores. Race differences for the CSJT are very small. In contrast, females do moderately better than males. This difference might be due to the confound between MOS and gender. For example, if we eliminate soldiers in the 11B MOS, the advantage for females is cut by more than 30%—to $d = .36$ and $d = .40$ for mini-forms *a* and *b*, respectively. Alternatively, the difference could be due to the typically better verbal skills of females compared to males.

Table 5.4. CSJT Scores by Gender

		Male		Female	
Score	d_{FM}	*M*	*SD*	*M*	*SD*
Mini-Form *a*	**0.66**	4.37	.53	4.53	.45
Mini-Form *b*	**0.51**	4.46	.60	4.77	.44

Note. $n_{Male(a,b)}$=131, 142. $n_{Female(a,b)}$=25, 22. d_{FM}= Effect size for Female-Male mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of the total group. Referent groups (e.g., Males) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

Table 5.5. CSJT Scores by Race/Ethnic Group

			White		Black		White Non-Hispanic		Hispanic	
Scores	d_{BW}	d_{HW}	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Mini-Form *a*	–.01	–.33	4.40	.57	4.40	.49	4.43	.54	4.26	.57
Mini-Form *b*	.07	–.11	4.47	.60	4.51	.52	4.47	.60	4.54	.51

Note. $n_{White(a,b)}$ = 101,99. $n_{Black(a,b)}$ = 23,26. $n_{White\ Non\text{-}Hispanic(a,b)}$ = 97,98. $n_{Hispanic(a,b)}$ = 22,23. d_{BW} = Effect size for Black-White mean difference. d_{HW} = Effect size for Hispanic-White Non-Hispanic mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., White) are listed second in the effect size subscript. The effect sizes are not statistically significant, $p < .05$ (two-tailed).

Table 5.6. CSJT Scores by MOS

				AW		11B		31U	
Variable	$d_{AW\text{-}11B}$	$d_{AW\text{-}31U}$	$d_{31U\text{-}11B}$	*M*	*SD*	*M*	*SD*	*M*	*SD*
Mini-Form *a*	**.55**	**–.44**	**1.00**	4.43	.53	4.14	.55	4.66	.34
Mini-Form *b*	.31	**–.43**	**.71**	4.50	.59	4.31	.65	4.75	.49

Note. n_{AW} = 156,165. n_{11B} = 61,60. n_{31U} = 12,16. AW = Army-Wide. 11B = Infantryman. 31U = Signal Support Systems Specialist. $d_{AW\text{-}11B}$ = Effect size for AW-11B mean difference. $d_{AW\text{-}31U}$ = Effect size for AW-31U mean difference. $d_{31U\text{-}11B}$ = Effect size for 31U-11B mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* across all Soldiers. Referent groups (e.g., 11B) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

Conclusions

We expect that the concurrent validation form of the CSJT will perform, psychometrically, much like its constituent mini-forms. The failure of the factor analysis to map items onto performance dimensions is neither surprising nor of particular concern. Items were categorized into performance dimensions not to produce scale scores but rather to ensure that the content of the CSJT items adequately covered the targeted performance domain. Furthermore, the test development process (i.e., generation of critical incidents and situations) ensured that the scenarios in the CSJT represent situations experienced by first-term Soldiers. Chapter 14 shows the relationships between the CSJT and several other measures. Among criterion measures, the CSJT correlates highest with teamwork, effort/initiative (which is conceptually related to self-management and self-directed learning), and supporting peers. This is evidence that the CSJT is tapping its targeted performance dimensions.

The items in the two mini forms comprise the items in the single CSJT form that will be used in the validation. Items that perform poorly (e.g., do not correlate with any other CSJT items) in the validation, however, might be dropped from the validation analyses.

CHAPTER 6: THE PERSONNEL FILE FORM AND ARCHIVAL CRITERION DATA

Dan J. Putka and Roy Campbell
HumRRO

Overview

In this chapter we discuss criterion data obtained from two sources: the Select21 Personnel File Form (S21-PFF) and the Army's Enlisted Master File (EMF). First we describe development of the S21-PFF and present results of its field test among current first-term Soldiers. Following presentation of the S21-PFF, we discuss two types of data being obtained from archival records, namely promotion rate and attrition status.

Select21 Personnel File Form

Description of Measure

The design of the S21-PFF is based largely upon the Project A and NCO21 Personnel File Forms (C. H. Campbell et al., 1990; Putka & R. C. Campbell, 2004). The content is drawn primarily from the Army's Promotion Point Worksheet (PPW; Department of the Army, 2000, May). The PPW serves as the basis for the Army's current semi-centralized NCO promotion system (Department of the Army, 2004, January). Soldiers receive promotion points in six areas: (a) Duty Performance Evaluation; (b) Promotion Board Points; (c) Awards, Decorations, and Achievements; (d) Military Education; (e) Civilian Education; and (f) Military Training. Promotion points for the first two areas are subjective, awarded by a Soldier's commander and promotion board members at the time a Soldier is up for promotion, whereas points for the latter four areas are objectively allocated by personnel administrators based on Soldiers' records. The S21-PFF contains sections that assess Soldiers' standing in the latter four areas of the PPW. Additionally, the S21-PFF has items on disciplinary actions (e.g., number of Article 15s), accelerated advancement, initial entry training (IET) performance, specialized skills, and Common Task Test (CTT) performance.

As noted above, administrative personnel typically complete most of the PPW based on Soldier records. The decision to use a self-report instrument to gather such data for Select21 rather than collecting the information through administrative means was based on positive experiences with the Project A and NCO21 PFF. In Project A, researchers found that archival records were not as current, nor as complete in comparison to self-report data provided by Soldiers (C. H. Campbell et al., 1990). For example, during pilot testing of the Project A PFF, researchers found that Soldiers reported higher numbers of both positive (e.g., Awards) and negative documents (e.g., Article 15s) compared to official records (Riegelhaupt, Harris, & Sadacca, 1987). This was explained by the fact that official records are not as current as self-report (even if updated frequently), and that some documents are removed from the official records after a certain amount of time. Furthermore, the fact that Soldiers reported more positive *and* negative documents, suggests that impression management typical of self-report measures may not be a concern here. Lastly, the self-report method provides the data substantially quicker and cheaper than is possible via administrative review of archival records. Given these considerations, there was good precedent to continue to gather such data via self-report in Select21.

Development of the S21-PFF

Because the design and content of the S21-PFF closely resemble PFFs used in past ARI projects, development of the S21-PFF was fairly straightforward. Project staff reviewed items appearing on past PFFs to make sure that they were still appropriate for the current Army (e.g., specific awards were still offered). The vast majority of items appearing on these PFFs were carried forward for use on the S21-PFF, with only minor revision. Next, we reviewed each PPW content area that was tapped by the PFF (e.g., Awards, Military Education), to determine if any new items should be added. Based on this review, two distinctions were added to the list of awards (Excellence in Armor, Excellence in Cavalry).[10]

After finalizing items for the PPW content areas, we brainstormed ways to expand upon the content of the PFF to broaden coverage of the criterion domain. Members of the Select21 Army Steering Committee (ASC) and Subject Matter Expert Panel (SMEP) had asked that we explore the use of existing Army tests for utilization as criteria. Project staff therefore generated a list of tests being used operationally that we might ask about on the PFF (e.g., Tank Crew Gunnery Skills Test [TCGST]; CTT). We evaluated the possibility of including questions regarding these tests on the S21-PFF based on four criteria: (a) standardization of the test across the Army, (b) eligibility considerations (not all Soldiers are eligible to take all tests), (c) verifiability of scores, and (d) feasibility of self-report assessment. Upon reviewing tests in light of these criteria, only questions about Soldiers' performance on the CTT were added to the S21-PFF. In later sections, we describe the CTT in more detail.

In addition to considering new content based on operational Army tests, we also explored other ways in which we could broaden the coverage of the S21-PFF. Through discussion among project staff and review of past research (e.g., Project A), we identified four additional areas for expansion: (a) disciplinary actions taken against the Soldier, (b) rate of promotion, (c) IET performance, and (d) additional skills or special qualifications that Soldiers have in their MOS. Items targeting each of these content areas were developed and added to the S21-PFF. Efforts to develop scales comprising items from each of these content areas are described in later sections.

Prior to field-testing the S21-PFF, we asked NCOs to carefully review the instrument. NCOs were asked to note any questions that seemed awkward, or would be unclear to Soldiers with 18-36 months of service. They were also asked to indicate whether any of the content on the S21-PFF was out of date or inaccurate (e.g., if certain awards are no longer offered). The NCOs indicated that no changes were needed.

Format of the S21-PFF

Each section of the PFF addresses one of the content areas described above (e.g., awards, military training). The S21-PFF employs a number of response formats that vary by content area examined. For example, some items use a checklist format and others require the entry of information (e.g., number of college credits). Additionally, the S21-PFF is computerized, which allows us to specify ranges of valid numeric responses to several questions on the measure (e.g.,

[10] These two distinctions are not given Army wide. They are only available to Soldiers in the 19K (M1 Armor Crewman) and 19D (Cavalry Scout) MOS.

number of certificates, fitness test scores). Imposing such constraints prevents respondents from making out of range responses, which in turn results in more useable data.

Field Test

Sample

Data on the S21-PFF were gathered from 318 Soldiers as part of the criterion field test. Information on the demographic characteristics of Soldiers who completed the S21-PFF, as well as other Select21 criterion measures, are provided in Chapter 1. Criterion field test problem logs revealed potential issues with four Soldiers' S21-PFF data. These Soldiers were excluded from all analyses in this chapter.

Scale Development

We attempted to create scales corresponding to each content area on the S21-PFF. Several of these scales reflected content and scoring algorithms used in past versions of the PFF, while other scales reflected new content for Select21. For new content areas, rational scoring algorithms were developed. For some content areas, the items were extremely heterogeneous, and did not warrant combining them into a single scale. In such cases we examined items from such areas separately as single item measures. Table 6.1 shows a listing of the scales we developed, and the single item measures that we examined. Below, we provide a description of our scale development efforts and each scale's final scoring.

Table 6.1. S21- PFF Scales and Single Item Measures

Scale	Single Item Measure
Awards	Additional Skill Identifier (ASI)
Military Education	Skill Qualification Identifier (SQI)
Army Physical Fitness Test (APFT)	IET- Exceptional Soldier Designation
Weapons Qualification	IET- Fast Track Program
Deviance	IET- Repeated Part of Training
Common Task Test (CTT) Attempts	Accelerated Advancement to E2
Simulated PPW	Accelerated Advancement to E3
	Accelerated Advancement to E4
	Promotion to E5 Waiver

Awards. The Awards scale is a weighted sum of awards (e.g., Purple Heart), military academic honors (e.g., Distinguished Honor Graduate), military board achievements (e.g., Soldier of the Quarter), certificates of achievement, and memoranda/letters of commendation earned by a Soldier. The S21-PFF asks Soldiers whether they received each type of distinction and, for some awards, the number of times they received each distinction. Each of these distinctions is worth a given number of promotion points for purposes of calculating a Soldier's promotion score on the PPW (Department of the Army, 2004, January).[11] These point assignments were used as weights to score Awards scale content.

[11] No formal promotion points are offered for memoranda/letters of commendation. However, given the similarity of their content to other distinctions in the Awards scale (e.g., certificates of achievement), we included memoranda/letters when calculating the Awards scale score. In calculating the Awards score, memoranda/letters were assigned a weight of 5 points which weights them comparably to Certificates of Achievement (5 points per certificate are offered on the PPW).

Military Education. Soldiers can earn promotion points for completing various military education courses, such as the Primary Leadership Development Courses (PLDC), Special Forces Qualification Course, Airborne School, and NBC School. As was the case with awards, completion of these military education courses are worth different numbers of points depending, in general, on their levels of prestige (Department of the Army, 2004, January). For example, the Special Forces Qualification Course is worth more points than Airborne School. For Select21, we created a Military Education scale by summing the promotion points associated with completion of various training courses.

Army Physical Fitness Test (APFT). The APFT consists of push-ups, sit-ups, and a 2-mile run. The APFT score is derived from a conversion table that takes raw test data (e.g., number of push-ups, time for the 2-mile run) and transforms it based on the age and gender of the Soldier (Department of the Army, 1998, October). The APFT scores reported here are based on conversion tables used for incorporating APFT data into the PPW (Department of the Army, 2004, January). These "rescaled" APFT scores range from 0 to 50.

Weapons Qualification. The S21-PFF asks Soldiers to indicate the last weapons qualification score they received on their individual weapon (e.g., M16 rifle, M4 carbine, M9 pistol). Weapons qualification scores are scaled on a metric used in past versions of the PPW (Unqualified = 0 points, Marksman = 10 points, Sharpshooter = 30 points, Expert = 50 points).[12] This same metric was used to generate weapons qualifications scores in NCO21.

Deviance. Five items on the S21-PFF ask Soldiers about their acts of misbehavior and disciplinary actions taken against them while in service. Two items ask Soldiers how many Flag Actions and Article 15s they have received. A Flag Action is a suspension of favorable personnel actions directed towards the Solder (e.g., preventing awards, reenlistment, and payment of bonuses to the Soldier). An Article 15 is a form of non-judicial punishment; specifically, it is a means for commanding officers to punish individuals for acts of deviant behavior. Soldiers are also presented with three "yes/no" items that ask if they were ever (a) court martialed, (b) arrested by civilian or military authorities while on active duty, or (c) given a written counseling statement.[13]

Although each of these items is indicative of misbehavior, the actions they ask about differ in their severity (e.g., a court martial is more serious than an Article 15). As such, we derived a rational weighting scheme for combining these items that allowed items reflecting more serious actions to receive more weight. Prior to deriving this weighting scheme, we examined base rates for the three "yes/no" items to evaluate whether they exhibited reasonable levels of variation. These analyses revealed that 4.8% of Soldiers were arrested by either civilian or military authorities while on active duty, 52.2% had received a written counseling statement, and no Soldiers had been court martialed. Based on these results, we excluded the court martial item from further consideration.

---

[12] A fairly recent change to the PPW resulted in a more complicated method for obtaining this score that factors in such factors as the type of weapon used and the type of targets engaged (Department of the Army, 2004, January). We used the original formula because of limitations in terms of the type of data we could collect via self-report.

[13] For scoring purposes, "yes" responses were coded as 1, and "no" responses were coded as zero.

Based on the severity of the actions they reflected, we determined that the Article 15 item should account for roughly 50% of the variance in the resulting Deviance scale score, Flag Actions for 20%, civilian/military arrests for 20%, and written counseling statements for 10% of the variance. With these percentages as targets, we used dominance analysis methods to assess the percentage of variance accounted for in the Deviance score by each item when simple unit weights were applied to them (Johnson, 2000). This initial analysis indicated that if we simply summed the items, the Article 15 item would account for 34.8% of the variance in the resulting Deviance score, Flag Actions for 40%, civilian/military arrests for 3.9%, and written counseling statements for 21.3% of the variance. Given these percentages diverged substantially from our targets, we iterated through a few alternative sets of weights until we arrived at percentages that were similar to our targets. The final set of weights, as well as the proportions of variance accounted for by the items once these weights were applied to them, are presented in Table 6.2. Results obtained with unit weights are provided for reference.

Table 6.2. Weighting of Items for the S21-PFF Deviance Scale

	Unit Weighting		Final Weighting	
Item	Weight	% Variance	Weight	% Variance
Article 15s	1	34.8	0.275	49.3
Flag Actions	1	40.0	0.100	21.0
Civilian/Military Arrests	1	3.9	0.525	20.6
Written Counseling Statements	1	21.3	0.100	9.1

Note. Weight = Weight applied to raw item score. % Variance = Percentage of variance in Deviance scale score attributable to each item.

Common Task Test (CTT) Attempts. The S21-PFF asks Soldiers about their performance on the 12 skill level 1 tasks that comprise the FY04 CTT (Manual for Administration of the FY04 Common Task Test, 2003). The purpose of the CTT is to provide commanders with a means for assessing Soldiers fundamental combat and survival skills (Department of the Army, 2003, April). Based on Soldiers' performance on the CTT, commanders can identify areas of weakness and take corrective action as needed. The tasks that comprise the CTT are not MOS-specific—they are designed to apply to all Soldiers. Normally, the CTT is administered once annually to all Soldiers.

There are several complications that may arise with trying to use the CTT as a criterion for Select21. First, the tasks that comprise it can shift from year to year. Each year tasks are selected based on input provided by a variety of Army commands. The field test version of the S21-PFF reflects the content of the FY03 CTT. For the concurrent validation effort, these tasks will likely need to be updated to reflect the FY04 CTT.[14] Another potential problem is that the Army has no formal requirements for reporting the results. Although Soldiers are provided with pass-fail information, numeric scores are not available. Lastly, Soldiers continue to take the CTT until they perform each task correctly. Thus, with no final numeric score, and every Soldier eventually passing the CTT, differentiating between Soldiers based on their performance on the

[14] Although tasks change from year to year, there is also considerable overlap; only about 3-5 tasks change each year.

CTT is difficult.[15] In light of these characteristics, the information we are able to acquire on the CTT via the S21-PFF is limited.

Given that all Soldiers eventually perform each task correctly, the final 12 questions on the S21-PFF ask Soldiers to indicate the number of attempts that it took them to receive a "GO" on each of the 12 tasks that comprise the FY04 CTT.[16] In responding to these questions, Soldiers are asked to use one of the following response options: 1, 2, 3, 4, "5 or more," or "Not sure." We calculated a CTT Attempts scale score for each Soldier by averaging across Soldiers' responses on each task (treating responses of "5 or more" as 5, and "Not Sure" as missing).[17] Using this metric, scores can range from 1 to 5, with lower scores indicating Soldiers successfully completed the CTT tasks in fewer tries. If Soldiers indicated they were "Not Sure" about their performance on more than three of the tasks (i.e., more than 25%), they were not given a CTT Attempts scale score.

The estimated internal-consistency reliability of the CTT Attempts scale was .92. Although the internal consistency reliability of the scale was high, there were several problems with the CTT data. First, 51 of the 314 Soldiers with PFF data (16.2%) indicated they were "not sure" of how many attempts it took them to pass at least four of CTT tasks, and as such, had no CTT Attempts score. Second, 164 of the 263 Soldiers with CTT Attempts scores (62.3%) had scores of "1," indicating that well over half the Soldiers passed all the CTT tasks on their first attempt. As we discuss later, such an extremely skewed "rare events" distribution may be problematic for analyses being planned for the concurrent validation effort. Despite these characteristics, we will examine the functioning of the CTT Attempts scale in this chapter.

Simulated PPW (SimPPW). In addition to the aforementioned scales, we calculated a composite score that simulates a Soldier's overall PPW score. The method we used to calculate this score was identical to the approach used in NCO21 (Putka & R. C. Campbell, 2004). This composite consists of the sum of four simulated PPW scale scores derived from S21-PFF content: Awards, Military Education, Civilian Education, and Military Training. In Select21, the SimPPW composite is being used as a criteria reflecting "promotability to the NCO level" that simulates what the Army currently uses to make promotions to the E5 and E6 pay grades. Although the maximum score that a Soldier could receive on this simulated composite is 500, the maximum score on the operational PPW is 800. The difference in point totals arises from the fact that the simulated PPW does not include Commander's evaluation points (max 150) or Promotion Board points (max 150).

Additional Skill Identifiers and Skill Qualification Identifiers. Questions regarding whether Soldiers have any additional skill identifiers (ASI) or special qualification identifiers (SQI) also appear on the S21-PFF. ASI are designations that identify Soldiers who have received specialized training in an area related to their MOS (e.g., sniper, pathfinder). Soldiers who have ASI have specialized training above and beyond the typical Soldier in their MOS. SQI are used

[15] This is even more confounded by the guidance that the preferred method for administering the CTT is in conjunction with the performance of "normal field training." In other words, the Soldier need not even know s/he is being evaluated (R. C. Campbell, personal communication).

[16] To receive a "GO" on a common task, Soldiers must perform all elements of it correctly.

[17] Treating responses of "5 or more" as "5" is based on the assumption that very few Soldiers will have failed the same task more than five times.

to identify Soldiers who meet requirements for certain assignments (e.g., parachutist, linguist, drill sergeant). SQI are more general designations than ASI in that they are not necessarily linked to specific MOS.

For the field test analyses, we examined the ASI and SQI items separately. Soldiers responses to these items were scored as 1 if the Soldier had an ASI/SQI, and 0 if the Soldier did not have an ASI/SQI. The rationale behind using ASI and SQI data as criteria is that they reflect specialized or advanced training. We examined base rates for each of these items to evaluate whether they exhibited reasonable levels of variation. These analyses revealed that 18.5% of Soldiers had an ASI, and 3.8% of Soldiers had an SQI.

Initial Entry Training (IET) Performance. Three items on the S21-PFF regard Soldiers' performance in IET. One question asks if Soldiers were in the top 10% of their unit in Basic Training or One Station Unit Training (OSUT). Another question asks if Soldiers were designated as part of the Fast Track Program (indicating high performance) in OSUT or Advanced Individual Training (AIT). Soldiers' responses to both of these questions were scored 1 if they answered "Yes," and 0 if they answered "No." A final question asks Soldiers if they had ever repeated a substantial part (over 24 hours) of their initial entry training. Soldiers' responses to this question were scored 1 if they if they answered "Yes," and 0 if they answered "No."

Our initial plan was to simply sum these three items to form an IET Performance scale. Prior to forming the scale, we examined base rates for each of these items to evaluate whether they exhibited reasonable levels of variation. These analyses revealed that 19.4% of Soldiers had the Exceptional Soldier Designation, 8.9% were in the Fast Track program, and 6.4% had repeated some part of training. Next we examined the internal consistency reliability (KR-20) that resulted from summing the three items, and item-deleted KR-20 statistics. These analyses indicated that the reliability of the scale was .32. Dropping the "repeated training" item would have raised the reliability to .47, however this was deemed inadequate for purposes of scale formation. Given the heterogeneity observed among these items, we decided to examine each IET performance item separately.

Accelerated Advancement. A series of three questions on the S21-PFF ask Soldiers if they received accelerated advancement to the E2, E3, or E4 pay grades. An additional question asks Soldiers if they received a waiver for promotion to the E5 pay grade. Soldiers' responses to each of these questions were scored 1 if they answered "Yes," and 0 if they answered "No." Like IET performance, our initial plan was to simply sum these four items to form an Accelerated Advancement scale. Prior to forming the scale, we examined base rates for each of these items to evaluate whether they exhibited reasonable levels of variation. These analyses revealed that 20.1% of Soldiers received an accelerated promotion to E2, 40.8% received an accelerated promotion to E3, 34.7% received an accelerated promotion to E4, and 12.4% received a waiver for promotion to E5. Next we examined the internal consistency reliability (KR-20) that resulted from summing the four items, and item-deleted KR-20 statistics. These analyses indicated that the reliability of the scale was only .33. Item-deleted KR-20 statistics indicated that dropping any of the items from the scale would not result in a notable increase in reliability. Given the heterogeneity observed among these items, we decided to examine each accelerated advancement item separately.

Results of Data Analysis

Because the S21-PFF was administered via computer, accurate data on administration times were available. The average time it took Soldiers to complete the S21-PFF was 11 minutes (*Mdn* = 7.6, *SD* = 12.8). 90% of Soldiers completed the PFF in less than 16 minutes, while 95% of Soldiers completed it in less than 25 minutes. This information will be useful as revisions to the S21-PFF are considered for the concurrent validation data collections.

Descriptive statistics. Table 6.3 shows descriptive statistics for each of the S21-PFF scales and single item measures described above. Examination of score distributions revealed several findings of note. Several scales exhibited moderate to high levels of positive skew, most notably Awards (*Skew* = 1.63), Military Education (*Skew* = 2.19), Deviance (*Skew* = 2.68), and CTT Attempts (*Skew* = 4.64). Graphical displays of the response distributions for these scales are presented in Figure 6.1.

Table 6.3. Descriptive Statistics for S21-PFF Scores

Variable	*n*	*Min*	*Max*	*M*	*SD*
Scale					
Awards	314	0	228	35.08	34.86
Military Education	314	0	96	10.05	13.31
Army Physical Fitness Test	298	0	50	25.78	11.51
Weapons Qualification	313	0	50	30.00	16.15
Deviance	314	0	1.95	0.18	0.28
CTT Attempts	263	1	3.55	1.13	0.33
Simulated PPW	314	10	366	106.21	51.27
Single Item Measure					
Additional Skill Identifier	314	0	1	0.18	0.39
Skill Qualification Identifier	314	0	1	0.04	0.19
IET- Exceptional Soldier Designation	314	0	1	0.19	0.40
IET- Fast Track Program	314	0	1	0.09	0.29
IET- Repeated Part of Training	312	0	1	0.06	0.25
Accelerated Advancement to E2	314	0	1	0.20	0.40
Accelerated Advancement to E3	314	0	1	0.41	0.49
Accelerated Advancement to E4	314	0	1	0.35	0.48
Promotion to E5 Waiver	314	0	1	0.12	0.33

Note. All single item measures were scored as follows: Yes = 1, No = 0.

As alluded to previously, such highly skewed distributions create difficulty for interpreting the results of inferential statistical tests based on assumptions of normality. Unfortunately, many of the analyses involved in validating predictors against these criteria that are being planned for the concurrent validation effort involve assumptions of normality (e.g., Pearson *r*, traditional multiple regression). To the extent that assumptions are violated, standard errors for statistics indexing the relationship between these scales and predictors may be biased downward, thus resulting in an inflated Type I error rate. To remedy this issue, we would need to assess relationships between predictor variables and these scales with statistical models specifically designed to deal with positively skewed criteria, such as the Poisson or negative binomial models discussed by Agresti (2002). Although using such models will address the issue

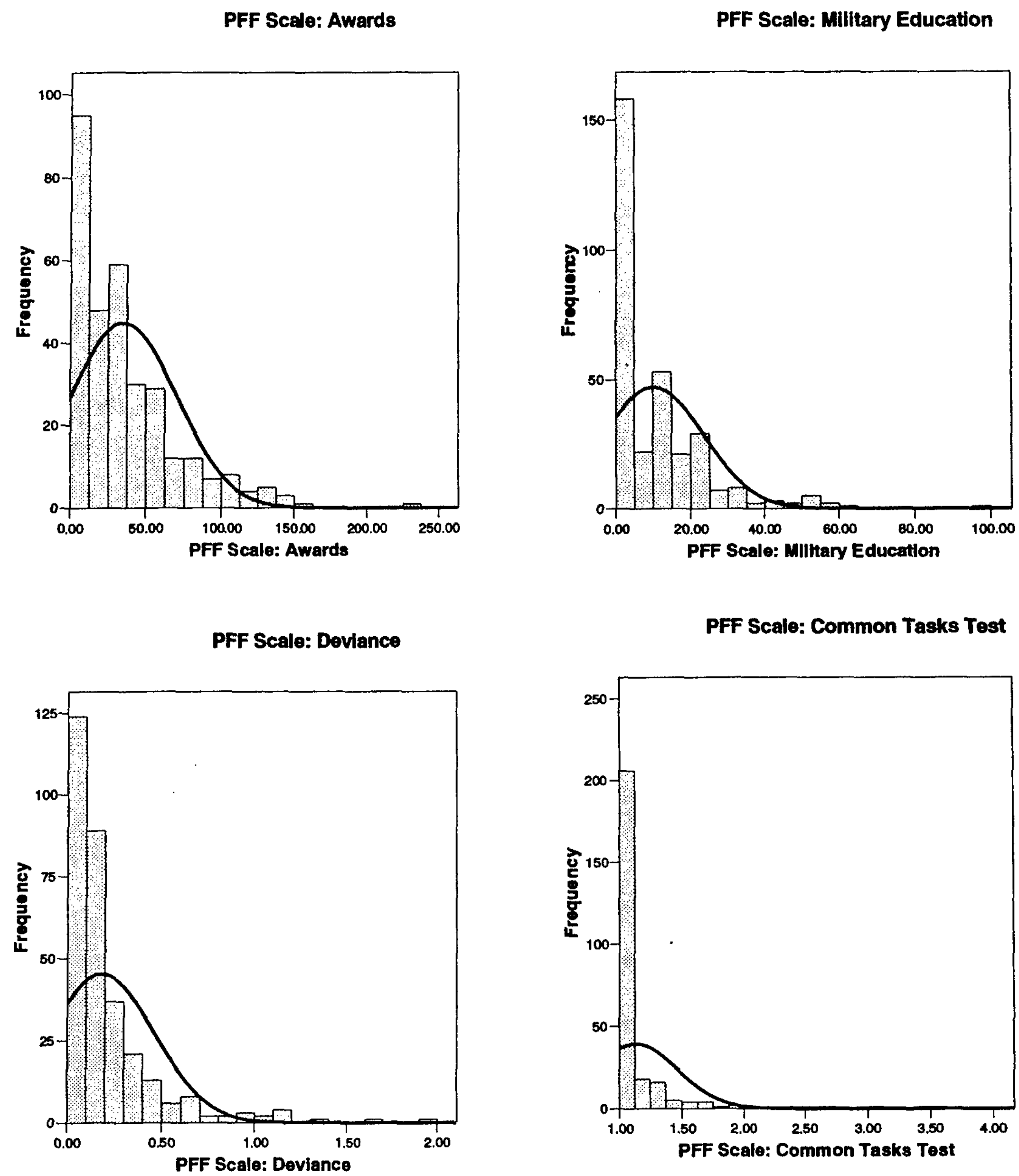

Figure 6.1. Response distributions for highly skewed S21-PFF scales.

of validating predictors against these scales, they would not necessarily help in cases where we were examining the underlying structure of the criterion domain (assuming these scales were included). In that case, methods for conducting exploratory or confirmatory factor analyses that are robust to violations of normality would need to be employed (Muthen & Muthen, 2001).

Scale intercorrelations. Table 6.4 shows intercorrelations among the S21-PFF scales and single item measures. Not surprisingly, the highest correlations are between the Simulated PPW and the S21-PFF scales that underlie it (i.e., Awards, Military Education, APFT, Weapons Qualification). Beyond the aforementioned set of correlations, no correlations exceeded .35 in magnitude, and most were substantially smaller than .20. Such results suggest that the scales and single item measures formed from S21-PFF content hold much unique variation.

Correlation of S21-PFF scores with time in service. Because some of the content on the S21-PFF may be a function of experience (e.g., number of awards), we examined correlations between Soldiers' days in service and their S21-PFF scores (see Table 6.5). The relationship between experience and S21-PFF scores was a concern because it could attenuate relationships between Select21 predictor measures and S21-PFF criteria. Days in service were significantly related to only three of the S21-PFF scores: Awards, Military Education, Simulated PPW, and Accelerated Advancement to E4. Although significant, these correlations were relatively small, with days in service accounting for only 3% - 5% of the variance in these scores.[18]

Subgroup differences on S21-PFF scores. Tables 6.6 through 6.8 show mean S21-PFF scores by gender, race/ethnicity, and MOS, respectively. With the exception of the Weapons Qualification scale, no sizable gender differences were found on the S21-PFF. Weapons Qualifications scores for men were an average of 0.72 standard deviations higher than such scores for women. Similarly, with regard to race/ethnicity, few sizeable differences were found. The exceptions were for Weapons Qualifications (Blacks scored 0.58 *SD*s lower than Whites), Simulated PPW (Blacks scored 0.48 *SD*s lower than Whites), and IET- Repeated Part of Training (Hispanics repeated a part of training more often than White Non-Hispanics, $d = 0.67$).

Lastly, several differences emerged for MOS. The majority of these differences were found on the Awards, Weapons Qualification and Simulated PPW scales. On the Awards scale, 11B Soldiers scored roughly 0.60 *SD*s higher than Soldiers in the 31U MOS. Interestingly, the mean Awards score for 11B Soldiers was not much different than the mean Awards score for Soldiers in Army-wide MOS ($d = -0.15$). Similarly, on the Weapons Qualification scale, Soldiers in the 11B MOS scored 0.66 *SD*s higher than Soldiers in the 31U MOS, and 0.49 *SD*s higher than Soldiers in Army-wide MOS. On the Simulated PPW scale, 11B Soldiers scored roughly 0.43 *SD*s higher than Soldiers in the 31U MOS, but not much differently from Soldiers in Army-wide MOS ($d = 0.04$). Soldiers in the 31U MOS had Simulated PPW scores that were roughly one-half *SD* lower than Soldiers in Army-wide MOS.

[18] It possible that the relationship between experience and these scales would be stronger in a less restricted sample. Recall, we targeted Soldiers with between 12 and 36 months experience for the criterion field test.

Table 6.4. Intercorrelations among S21-PFF Scores

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Scale															
1 Awards	-														
2 Military Education	**.25**	-													
3 Army Physical Fitness Test	**.12**	.10	-												
4 Weapons Qualification	**.17**	.06	**.19**	-											
5 Deviance	.02	.01	-.09	.04	-										
6 CTT Attempts	.01	.03	.04	**-.18**	.03	-									
7 Simulated PPW	**.69**	**.50**	**.42**	**.44**	-.01	-.03	-								
Single Item Measure															
8 Additional Skill Identifier	**.14**	**.16**	.06	**.11**	-.03	-.07	**.15**	-							
9 Skill Qualification Identifier	.03	**.21**	.04	.08	.11	-.02	**.12**	**.20**	-						
10 IET- Exceptional Soldier Designation	**.18**	**.12**	**.19**	**.22**	.09	.01	**.27**	.08	**.11**	-					
11 IET- Fast Track Program	**.15**	**.15**	**.15**	.04	**.12**	-.01	**.16**	.08	**.11**	**.33**	-				
12 IET- Repeated Part of Training	-.05	.07	-.10	**-.15**	**.16**	.08	-.10	.08	.02	.07	-.04	-			
13 Accelerated Advancement to E2	.05	.00	.08	.03	.05	**.14**	.07	.01	-.02	.06	.04	-.03	-		
14 Accelerated Advancement to E3	-.06	.03	.09	.05	-.03	.07	.03	-.03	.00	.07	**.13**	-.09	**.33**	-	
15 Accelerated Advancement to E4	**.13**	**.11**	**.18**	.06	.04	**.15**	**.20**	-.02	.10	.03	-.02	.06	**.17**	.10	-
16 Promotion to E5 Waiver	**.26**	**.24**	.10	.08	.02	-.07	**.28**	-.01	.03	**.13**	**.15**	-.02	-.02	**-.12**	**.13**

Note. $n = 250\text{-}314$. Statistically significant correlations are bolded, $p < .05$ (two-tailed).

Table 6.5. Correlations between S21-PFF Scores and Days in Service

Variable	r
Scale	
Awards	**.21**
Military Education	**.17**
Army Physical Fitness Test	-.03
Weapons Qualification	.02
Deviance	.06
CTT Attempts	.04
Simulated PPW	**.21**
Single Item Measure	
Additional Skill Identifier	-.09
Skill Qualification Identifier	-.01
IET- Exceptional Soldier Designation	-.08
IET- Fast Track Program	.01
IET- Repeated Part of Training	.00
Accelerated Advancement to E2	.08
Accelerated Advancement to E3	-.03
Accelerated Advancement to E4	**.24**
Promotion to E5 Waiver	.05

Note. n = 262-313. Statistically significant are bolded, $p < .05$ (one-tailed).

Table 6.6. S21-PFF Scores by Gender

		Male		Female	
Variable	d_{FM}	*M*	*SD*	*M*	*SD*
Scale					
Awards	-0.23	36.44	35.80	28.15	28.33
Military Education	-0.16	10.40	13.28	8.30	13.54
Army Physical Fitness Test	-0.22	26.17	11.59	23.67	11.01
Weapons Qualification	**-0.72**	31.81	16.21	20.21	11.70
Deviance	**-0.27**	0.19	0.29	0.12	0.15
CTT Attempts	-0.05	1.13	0.35	1.12	0.21
Simulated PPW	-0.30	108.70	49.59	93.70	58.35
Single Item Measure					
Additional Skill Identifier	-0.23	0.20	0.40	0.11	0.31
Skill Qualification Identifier	**-0.22**	0.05	0.21	0.00	0.00
IET- Exceptional Soldier Designation	**-0.31**	0.21	0.41	0.09	0.28
IET- Fast Track Program	0.07	0.09	0.28	0.11	0.31
IET- Repeated Part of Training	0.10	0.06	0.24	0.09	0.28
Accelerated Advancement to E2	0.03	0.20	0.40	0.21	0.41
Accelerated Advancement to E3	-0.05	0.41	0.49	0.38	0.49
Accelerated Advancement to E4	-0.02	0.35	0.48	0.34	0.48
Promotion to E5 Waiver	-0.21	0.14	0.34	0.06	0.25

Note. n_{Male} = 223-266, n_{Female} = 39-47. d_{FM} = Effect size for Female-Male mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., Males) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

Table 6.7. S21-PFF Scores by Race/Ethnic Group

			White		Black		White Non-Hispanic		Hispanic	
Variable	d_{BW}	d_{HW}	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Scale										
Awards	-0.24	-0.13	38.04	37.40	29.06	30.83	37.76	37.00	33.04	33.38
Military Education	**-0.37**	**-0.32**	11.32	13.62	6.35	12.66	11.56	13.80	7.08	9.83
Army Physical Fitness Test	-0.30	0.02	26.37	11.36	22.96	12.39	26.27	11.34	26.54	11.89
Weapons Qualification	**-0.58**	-0.11	32.47	15.95	23.27	15.43	32.20	16.24	30.42	15.70
Deviance	0.32	0.12	0.17	0.26	0.25	0.34	0.17	0.26	0.20	0.34
CTT Attempts	0.36	0.06	1.10	0.30	1.21	0.45	1.12	0.36	1.14	0.31
Simulated PPW	**-0.48**	-0.17	113.04	53.96	87.25	44.00	113.01	55.17	103.42	44.07
Single Item Measure										
Additional Skill Identifier	**-0.28**	0.08	0.21	0.41	0.10	0.30	0.20	0.40	0.23	0.42
Skill Qualification Identifier	**-0.22**	0.07	0.05	0.21	0.00	0.00	0.05	0.21	0.06	0.24
IET- Exceptional Soldier Designation	0.26	0.26	0.17	0.38	0.27	0.45	0.17	0.38	0.27	0.45
IET- Fast Track Program	0.06	0.27	0.08	0.27	0.10	0.30	0.07	0.26	0.15	0.36
IET- Repeated Part of Training	0.15	**0.67**	0.05	0.21	0.08	0.27	0.04	0.19	0.17	0.38
Accelerated Advancement to E2	-0.15	0.05	0.22	0.41	0.15	0.36	0.21	0.41	0.23	0.42
Accelerated Advancement to E3	0.05	0.02	0.38	0.49	0.40	0.50	0.39	0.49	0.40	0.49
Accelerated Advancement to E4	-0.08	0.22	0.35	0.48	0.31	0.47	0.35	0.48	0.46	0.50
Promotion to E5 Waiver	-0.10	-0.01	0.13	0.34	0.10	0.30	0.13	0.34	0.13	0.33

Note. n_{White} = 170-199. n_{Black} = 43-52. $n_{White\ Non\text{-}Hispanic}$ = 160-187. $n_{Hispanic}$ = 39-48. d_{BW} = Effect size for Black-White mean difference. d_{HW} = Effect size for Hispanic-White Non-Hispanic mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., White) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

Table 6.8. S21-PFF Scores by MOS

Variable	$d_{AW\text{-}11B}$	$d_{AW\text{-}31U}$	$d_{31U\text{-}11B}$	AW		11B		31U	
				M	*SD*	*M*	*SD*	*M*	*SD*
Scale									
Awards	-0.15	**0.43**	**-0.58**	37.75	34.27	42.80	40.52	22.59	25.99
Military Education	0.25	0.13	0.12	11.14	13.13	7.80	12.63	9.37	11.13
Army Physical Fitness Test	-0.13	-0.03	-0.10	25.35	11.73	26.88	11.91	25.70	8.29
Weapons Qualification	**-0.49**	0.18	**-0.66**	29.18	15.69	37.03	15.87	26.30	13.63
Deviance	-0.06	0.21	-0.26	0.20	0.27	0.21	0.32	0.14	0.21
CTT Attempts	-0.18	-0.21	0.03	1.10	0.19	1.16	0.42	1.17	0.42
Simulated PPW	0.04	**0.48**	**-0.43**	112.91	53.67	110.63	51.42	88.37	41.21
Single Item Measure									
Additional Skill Identifier	-0.22	-0.34	0.12	0.13	0.33	0.21	0.41	0.26	0.45
Skill Qualification Identifier	-0.04	0.00	-0.04	0.04	0.19	0.04	0.21	0.04	0.19
IET- Exceptional Soldier Designation	-0.08	0.25	-0.33	0.21	0.41	0.24	0.43	0.11	0.32
IET- Fast Track Program	-0.06	0.06	-0.12	0.09	0.29	0.11	0.31	0.07	0.27
IET- Repeated Part of Training	0.07	-0.01	0.08	0.07	0.26	0.05	0.23	0.07	0.27
Accelerated Advancement to E2	0.03	0.24	-0.21	0.21	0.41	0.20	0.40	0.11	0.32
Accelerated Advancement to E3	0.23	0.02	0.21	0.45	0.50	0.34	0.48	0.44	0.51
Accelerated Advancement to E4	0.01	0.06	-0.05	0.36	0.48	0.36	0.48	0.33	0.48
Promotion to E5 Waiver	-0.10	0.27	**-0.37**	0.13	0.33	0.16	0.37	0.04	0.19

Note. n_{AW} = 98-110. n_{11B} = 87-112. n_{31U} = 20-27. AW = Army-Wide. 11B = Infantryman. 31U = Signal Support Systems Specialist. $d_{AW\text{-}11B}$ = Effect size for AW-11B mean difference. $d_{AW\text{-}31U}$ = Effect size for AW-31U mean difference. $d_{31U\text{-}11B}$ = Effect size for 31U-11B mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* across all Soldiers. Referent groups (e.g., 11B) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

Discussion

In preparation for the concurrent validation effort, we will shorten the S21-PFF in light of results presented above. For the concurrent validation effort, we will retain all of the scales that contribute to the Simulated PPW (i.e., Awards, Military Education, Weapons Qualification, and APFT). Nevertheless, one change we will make to this content for the validation effort regards the APFT. On the current S21-PFF, Soldiers who failed the APFT did not provide scores. In the concurrent validation effort, we will ask Soldiers to report their latest APFT score regardless of whether they passed or failed. Although failing APFT scores are treated as 0 for purposes of contributing to the PPW, it is better not to restrict their range when APFT scores are used as a criterion.

Much new content was developed for the S21-PFF and piloted as part of the field test. At this point, we are recommending items regarding the CTT be dropped from the concurrent validation version of the S21-PFF because (a) a notable percentage of Soldiers indicated they were "not sure" of how many times it took them to pass tasks on the CTT, and (b) over 60% of the Soldiers had the same CTT score (i.e., Soldier who passed all tasks on the first try). Such a high level of missing data and skewness in resulting scores suggests that the CTT scale would be of limited utility as a criterion for Select21. If the decision is made to keep CTT items on the S21-PFF, it will be necessary to revise their content so that they reflect the FY05 CTT.

In addition to dropping CTT items from the S21-PFF, we also recommend dropping items regarding SQI and whether or not Soldiers had been court martialed due to lack of variation. Very few Soldiers indicated they had an SQI (3.8%), and no Soldiers indicated they had been court martialed.

Archival Criterion Data

In addition to collecting self-report data from Soldiers via the S21-PFF, we are also obtaining archival data from the EMF that allows for calculation of Soldiers' rate of promotion and attrition status. Here we provide a brief overview of plans for obtaining attrition data. Following discussion of the attrition data, we describe data gathered to calculate promotion rates, and present results of analyses examining promotion rates among Soldiers participating in the criterion field test.

Attrition Data

We are currently obtaining attrition data for Active Duty Soldiers who participated in the predictor field test, faking research, and pilot test phases of Select21. During these data collections, new recruits completed the Select21 predictor measures as they processed through their reception battalions (i.e., immediately upon entering service). Taken together, these data collections can be viewed as a mini-longitudinal validation sample where the primary criteria of interest are attrition through various months of service (e.g., through 6 months or through 12 months). Although the Select21 predictor measures were at various stages of development during these data collections, we appear to have sufficient data and continuity in measures across administrations to provide a reasonable assessment of how these measures fare in terms of predicting various types of attrition.

The plan for constructing a database of attrition information on participants in these data collections, as well as conducting and documenting attrition-related analyses in Select21, is detailed in a memorandum for record (Putka, 2004). Basically, the plan involves producing "on-demand" attrition reports that include, among other things, correlations between predictor scale scores and the most recent attrition data available at the time a report is requested. Such on-demand attrition reports are designed to support key project meetings (e.g., meetings with the Army Steering Committee or Scientific Review Panel). The reports will be designed to provide project stakeholders an update on how the Select21 predictor measures are faring with regard to predicting attrition. Towards the end of the project, more comprehensive attrition analyses will be undertaken in support of the final concurrent validation and recommendations report. Among other things, these final analyses will include a more comprehensive examination of relations between the person-environment fit measures and attrition criteria.

Promotion Rate

For Soldiers in the criterion field test, we obtained EMF data regarding their date of entry in to the Army and current pay grade. Data on these Soldiers' pay grade at entry was obtained from the U.S. Military Entrance Processing Command Integrated Resource System (MIRS) database. These data, along with Soldiers' testing date and MOS (obtained during the field test), allow for calculation of Soldiers' individual rates of promotion.

Sample

Promotion rates were calculated for 218 of the 339 Soldiers (64.3%) who completed measures as part of the criterion field test. Among Soldiers for whom no promotion rates were calculated, 49 were prior service accessions (14.5% overall), and 121 lacked pay grade at entry data (35.7%).[19] Current pay grades of Soldiers with promotion data ranged from 1 to 5 (M = 3.55, SD = 0.64), and their pay grades at entry ranged from E1 to E4 (M = 1.90, SD = 0.98).

Calculating Promotion Rates

In Project A, Soldiers' rates of promotion were indexed by residuals resulting from regression analyses conducted within each MOS (J. P. Campbell & Knapp, 2001, p. 220). Specifically, within each MOS, Soldiers' current pay grade was regressed on their time in service.[20] Positive residuals indicated that Soldiers advanced quicker than the average Soldier in their MOS with their time in service. Negative residuals indicated that Soldiers advanced slower than the average Soldier in their MOS with their time in service. An important feature of this method of calculating promotion rates is that it controls for differences in promotion rates across MOS. A large factor in determining promotions is vacancies in higher positions. The extent of such vacancies varies greatly by MOS. As such, taking MOS into account when calculating individual rates of promotion is critical. If MOS differences are not accounted for,

[19] Pay grade at entry was necessary for calculation of promotion rates; we were unable to retrieve this information for reserve component Soldiers.
[20] Time in service squared was also included as a predictor.

then individual differences in "promotability" may be obscured by factors that are beyond Soldiers' control (e.g., characteristics of a Soldiers' MOS).

One potential drawback of the method used in Project A is that it does not account for differences in Soldier pay grades at entry. For example, if it is easier and requires less time for Soldiers to advance from E1 to E3 than from E3 to E5, the fact that a Soldier who entered at E1 advanced faster than a Soldier who entered at E3 may have little to do with individual differences in the effectiveness of these two Soldiers (Riegelhaupt et al., 1987). Thus, although the Project A promotion rate score provides an accurate reflection of Soldiers' *actual* rate of promotion, it may not necessarily provide an accurate reflection of Soldiers' performance in service that lead to promotion. To remedy this, an alternative method for calculating the promotion score would be to add pay grade at entry as an additional predictor in the MOS-specific regression equations used in Project A. Doing so would essentially control for differences in Soldiers' pay grade at entry.

Unfortunately, unlike Project A, the number of Soldiers populating each MOS in the criterion field test sample was not sufficient to perform within-MOS regressions. Table 6.9 shows the number of Soldiers in each MOS for whom promotion rate data were available. As shown in Table 6.9, the majority of MOS samples had fewer than 10 Soldiers in them, and many MOS had just one Soldier. Thus, the within-MOS regression approach to calculating promotion rates was not feasible for analysis of the Select21 field test data.

Table 6.9. Number of Soldiers with Promotion Rate Data by MOS

MOS	*n*	MOS	*n*	MOS	*n*	MOS	*n*
11B	61	42L	8	19D	1	77F	1
11C	11	46Q	1	19K	4	82S	1
13B	8	52D	1	21B	2	91P	1
13D	5	56M	1	21T	1	91W	1
13E	1	63W	1	31B	5	92A	3
13F	2	73C	1	31C	5	92G	2
13S	1	73D	2	31F	2	92Y	2
14J	1	74B	28	31U	24	96B	23
15Q	1	74D	4	42A	2		

Note. Total $n = 218$. This sample included 35 different MOS.

Despite the small sample sizes within many of the MOS, we sought a way to incorporate MOS into our promotion rate calculations. For purposes of field test analyses, we calculated promotion rate three different ways, each of which has strengths and weaknesses. The first promotion rate score we calculated was based on residuals obtained from regressing Soldiers' current pay grade on their days in service, pay grade at entry, pay grade at entry squared, and MOS.[21] MOS was entered in this model as a series of 34 dummy variables.[22] A strength of this index is that it incorporates, time in service, pay grade at entry, and MOS into the promotion rate calculation. A limitation of this index is that the promotion rate score for Soldiers who are the sole representative of their MOS is 0. In other words, these Soldiers were

[21] Pay grade at entry squared was entered as an additional variable because it significantly incremented prediction of current pay grade.
[22] There was one dummy variable for each MOS except 11B, which served as the referent group.

treated as if they had advanced at the same rate as the "average" Soldier in their MOS, because no other Soldiers were available in their MOS to indicate whether they had advanced at a faster or slower rate.

Given the aforementioned limitation, we also generated a regression-based promotion rate score that excluded MOS as a variable. Specifically, the second promotion rate score we calculated reflected residuals obtained from regressing Soldiers' current pay grade on their days in service, pay grade at entry, and pay grade at entry squared. To assess the degree to which our ability to predict current grade was lost by omitting MOS, we examined the change in R^2 associated with eliminating the MOS dummy variables. Results for a series of models we fitted to predict Soldiers' current pay grade are presented in Table 6.10.

Table 6.10. Models of Current Pay Grade

Model	*Df*	R^2	ΔR^2
1. Days in Service	1	**.35**	**.35**
2. Days in Service + Pay Grade at Entry	2	**.40**	**.05**
3. Days in Service + Pay Grade at Entry (L,Q)	3	**.44**	**.04**
4. Days in Service + Pay Grade at Entry (L,Q) + MOS	37	**.59**	**.15**

Note. *Df* = Degrees of freedom for model. ΔR^2 = Change in R^2 from previous model in table. In Models 3 and 4, "(L,Q)" indicates that both a linear and quadratic (squared) term were included for pay grade at entry. Statistically significant R^2 are bolded, $p < .05$ (two-tailed).

The model underlying our first promotion rate score (Model 4) accounted for 59% of the variance in Soldiers' current pay grade. Eliminating MOS resulted in a model that accounted for 44% of the variance in Soldiers' current pay grade (Model 3). The significant change in R^2 between Models 3 and 4 suggests that MOS does appear to play an important role in predicting Soldiers' current grade. Nevertheless, although the drop in R^2 was substantial, Model 3 still accounted for a substantial amount variance in current pay grade. Given the issues associated with incorporating MOS into the equation noted above, using the residuals based on Model 3 as an index of promotion rate may be more meaningful for Select21.

Lastly, we generated a simple promotion rate score that reflected the number of pay grades Soldiers advance per year. This score was calculated by subtracting a Soldier's pay grade at entry from his/her current pay grade and dividing it by the number of years s/he had been in service. The strength of this score lies in its simplicity and interpretability. Its weaknesses are that it treats movement between any two pay grades as equally easy to achieve and ignores MOS.

Results of Analyses

Descriptive statistics and intercorrelations for the three promotion rate scores are presented in Table 6.11. Of note in Table 6.11 is the high degree of relationship between the two pay grade residual scores. Examination of these scores' distributions revealed that the pay grade residuals were fairly normally distributed, and the grades advanced per year score was more bimodal in nature with peaks centered around 0 and 1.04.

Table 6.11. Descriptive Statistics and Intercorrelations for Promotion Rate Scores

Variable	Descriptives				Correlations	
	Min	*Max*	*M*	*SD*	1	2
1 Pay Grade Residual w/ MOS	-1.47	1.4133	0.00	0.41	-	
2 Pay Grade Residual w/o MOS	-2.92	1.3212	0.00	0.48	0.80	-
3 Grades Advanced Per Year	-0.76	3.7655	0.84	0.58	0.40	0.42

Note. $n = 218$. All correlations were statistically significant ($p < .05$, one-tailed).

Correlations with S21-PFF scores. Table 6.12 shows correlations between promotion rate and S21-PFF scores. Interestingly, none of the promotion rate scores had significant positive correlations with the Simulated PPW score. These findings appear to cast doubt on the viability of the Simulated PPW as an index of "promotability" during Soldiers' first term. Given that the PPW is used for promoting Soldiers from E4 to E5, and from E5 to E6 pay grades, perhaps it is not surprising that it shows little correlation with rate of promotion within the E1 to E4 pay grade range—which constitutes the majority of the field test sample.

The S21-PFF scores with strongest relationships with promotion rate were accelerated advancement to E4 and Deviance. All three promotion rate scores were significantly positively related to accelerated advancement to E4. Interestingly, no significant relationships were found between the promotion rate scores and the other accelerated advancement items. Deviance was significantly negatively related to both pay grade residual scores. Such findings are consistent with expectations—Soldiers who have discipline problems tend to advance at lower rates relative to comparable Soldiers (i.e., in terms of days in service, pay grade at entry, and MOS).

Subgroup differences. Tables 6.13 through 6.15 show mean promotion rate scores by gender, race/ethnicity, and MOS, respectively. Examination of promotion rate scores by gender revealed no differences. With regard to race/ethnicity, Hispanics scored significantly higher than White Non-Hispanics on Pay Grade Residual with MOS ($d = 0.42$) and Grades Advanced Per Year ($d = 0.54$). No Black-White differences were observed. With regard to MOS, no notable differences were found. Interestingly, minimal differences were found between MOS in terms of the Pay Grade Residual without MOS score ($d = 0.01$ to 0.03). These latter findings suggest that rates of promotion in the MOS explicitly targeted by Select21 (i.e., 11B, 31U) did not substantially differ. Such results suggest that factoring MOS into the promotion rate calculation for Soldiers across these MOS may not be critical.

Discussion

The method we will use to calculate promotion rates scores for the concurrent validation effort will likely depend on the sampling plan adopted. Based on the pattern of correlations with the S21-PFF scores, it is likely that one of the two pay grade residual scores will be used as the final promotion rate criterion. To the extent that the majority of the validation sample is comprised of few, heavily sampled MOS (the more likely scenario), it might be best to adopt the pay grade residual score that incorporates MOS into its calculation. To the extent that the majority of the validation sample is comprised of many, scarcely sampled MOS, it might be best to adopt the pay grade residual score that excludes MOS from its calculation. Analysis of the field test data did not provide a clear choice regarding which method is best. Analyses indicated that both of these scores are highly related ($r = .80$), and both show similar levels of relationship

with other variables. However, as alluded to previously, these results were based on a sample that is heavily comprised of Soldiers in the target Select21 MOS. Differences in promotion rates for Soldiers in these MOS appeared to be relatively small. To the extent that the concurrent validation sample reflects broader sampling of Soldiers from several MOS, the similarity of these two scores would likely be lessened.

Table 6.12. Correlations between Promotion Rate and S21-PFF Scores

Variable	Pay Grade Residual (w/ MOS)	Pay Grade Residual (w/o MOS)	Grades Advanced Per Year
Scale			
Awards	-0.02	-0.07	-0.08
Military Education	0.06	0.09	-0.11
Army Physical Fitness Test	0.11	0.11	-0.09
Weapons Qualification	0.09	0.05	0.02
Deviance	**-0.25**	**-0.34**	-0.10
CTT Attempts	-0.03	-0.03	0.00
Simulated PPW	0.11	0.03	**-0.22**
Single Item Measure			
Additional Skill Identifier	0.01	0.02	-0.02
Skill Qualification Identifier	**0.16**	**0.16**	0.10
IET- Exceptional Soldier Designation	0.00	-0.04	**-0.12**
IET- Fast Track Program	-0.10	**-0.16**	**-0.14**
IET- Repeated Part of Training	0.02	0.01	0.03
Accelerated Advancement to E2	0.07	0.00	0.08
Accelerated Advancement to E3	0.06	0.00	0.11
Accelerated Advancement to E4	**0.36**	**0.37**	**0.19**
Promotion to E5 Waiver	0.10	0.07	-0.10

Note. n = 174-205. Statistically significant are bolded, $p < .05$ (one-tailed).

Table 6.13. Promotion Rate Scores by Gender

		Male		Female	
Variable	d_{FM}	*M*	*SD*	*M*	*SD*
Pay Grade Residual w/ MOS	-0.13	0.01	0.42	-0.05	0.41
Pay Grade Residual w/o MOS	0.01	0.00	0.49	0.00	0.42
Grades Advanced Per Year	-0.05	0.84	0.59	0.81	0.50

Note. n_{Male} = 186, n_{Female} = 31. d_{FM} = Effect size for Female-Male mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., Males) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

Table 6.14. Promotion Rate Scores by Race/Ethnic Group

Variable	d_{BW}	d_{HW}	White *M*	White *SD*	Black *M*	Black *SD*	White Non-Hispanic *M*	White Non-Hispanic *SD*	Hispanic *M*	Hispanic *SD*
Pay Grade Residual w/ MOS	0.15	**0.42**	-0.03	0.41	0.03	0.38	-0.05	0.41	0.13	0.40
Pay Grade Residual w/o MOS	-0.02	0.34	-0.02	0.44	-0.03	0.64	-0.03	0.44	0.12	0.43
Grades Advanced Per Year	0.21	**0.54**	0.78	0.51	0.89	0.63	0.75	0.51	1.03	0.69

Note. n_{White} = 143. n_{Black} = 35. $n_{White\ Non\text{-}Hispanic}$ = 131. $n_{Hispanic}$ = 37. d_{BW} = Effect size for Black-White mean difference. d_{HW} = Effect size for Hispanic-White Non-Hispanic mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/SD of referent group. Referent groups (e.g., White) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

Table 6.15. Promotion Rate Scores by MOS

Variable	$d_{AW\text{-}11B}$	$d_{AW\text{-}31U}$	$d_{31U\text{-}11B}$	AW *M*	AW *SD*	11B *M*	11B *SD*	31U *M*	31U *SD*
Pay Grade Residual w/ MOS	0.00	0.00	0.00	0.00	0.33	0.00	0.48	0.00	0.49
Pay Grade Residual w/o MOS	0.03	0.02	0.01	-0.01	0.54	-0.03	0.48	-0.02	0.50
Grades Advanced Per Year	0.00	-0.27	0.28	0.83	0.62	0.82	0.62	0.98	0.55

Note. $n_{Army\text{-}Wide}$ = 71. n_{11B} = 72. n_{31U} = 24. AW = Army-Wide. 11B = Infantryman. 31U = Signal Support Systems Specialist. $d_{AW\text{-}11B}$ = Effect size for AW-11B mean difference. $d_{AW\text{-}31U}$ = Effect size for AW-31U mean difference. $d_{31U\text{-}11B}$ = Effect size for 31U-11B mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* across all Soldiers. Referent groups (e.g., 11B) are listed second in the effect size subscript. No effect sizes were statistically significant ($p < .05$, two-tailed).

CHAPTER 7: ATTITUDINAL CRITERIA

Chad H. Van Iddekinge, Dan J. Putka, and Christopher E. Sager
HumRRO

Background

This chapter focuses on the attitudinal criterion measures developed for this project. Although various Select21 predictors may be related to attitudinal criteria (e.g., job satisfaction), they were developed primarily to validate the person-environment (P-E) fit predictor measures described in Chapter 13. The major criterion we would like the fit measures to predict is attrition. However, the concurrent design of the Select21 validation effort will not allow us to examine relations between the fit predictors and attrition in our primary sample. Therefore, we developed measures of constructs that theory and empirical evidence suggest are the strongest precursors of attrition. These measures assess attitudes and intentions that (a) have highly developed research literatures, (b) have theoretical and empirical relationships with attrition and predictor measures that could be used in a selection context, and (c) can be assessed in a concurrent validation design. The criteria we chose can be grouped into current-state and future-oriented criteria. Current-state criteria reflect Soldiers' current standing on a construct (e.g., current level of job satisfaction), whereas future-oriented criteria reflect Soldiers' expected future standing on a construct given anticipated future Army conditions. We describe results for each set of criteria in turn.

Army Life Survey

Description of Measure

Current-state attitudes/intentions are assessed in the Army Life Survey (ALS). The ALS is a 105-item instrument comprising 16 scales. These scales were developed based on a review of the relevant literatures, including research from the applied psychology literature (e.g., Hom & Griffeth, 1995; Jex, 1998; Meyer & Allen, 1997; Spector, 1997) and previous Army research, such as Project A (J. P. Campbell & Knapp, 2001) and Project First Term (Strickland, 2004). In fact, most of the ALS scales were adapted from established measures within the literature.

The 16 ALS scales can be grouped into three broad categories of criteria. The first category includes three scales that measure intentions to remain in the Army, including attrition intentions, re-enlistment intentions, and intentions to make the Army a career. Intentions are thought to be the strongest and most proximal predictor of actual behavior (Ajzen, 1991). Indeed, prior research has shown that intentions to quit are the best predictor of turnover in civilian jobs (e.g., Griffeth, Hom, & Gaertner, 2000) and attrition in the military (e.g., Strickland, 2004). The three ALS intentions scales were adapted primarily from measures developed for Project First Term, but also from existing measures within the civilian literature (e.g., Hom & Griffeth, 1995).

The second category of ALS criteria includes measures of several attitudinal variables that have been shown to underlie both intentions to leave and actual withdrawal behavior (e.g., Griffeth et al., 2000). These include satisfaction with various aspects of Army life (supervision, peers, work itself, promotions, pay and benefits, and the Army in general), organizational

commitment (affective, continuance, and normative commitment), perceived fit (with MOS and the Army in general), and perceived stress. The satisfaction scales were adapted from the Army Job Satisfaction Questionnaire developed in Project A. The commitment scales were based on Meyer and Allen's (1997) three-component model of commitment and were adapted from a widely used existing measure of these constructs (Meyer, Allen, & Smith, 1993). The perceived fit scales were adapted to the Army context using various measures available in the P-E fit literature (e.g., Cable & Judge, 1996). Finally, although the perceived stress scale was developed from a review of the stress literature (e.g., Jex, 1998), the items were not based on any specific existing measure.

The final category of ALS criteria consists of a single scale in which respondents are asked to rate how important they believe the core Army values are to being an effective Soldier. This scale was included because Soldiers who fit well in the Army may be more likely to endorse its core values. Indeed, past research has linked such values to attrition within the Army (Strickland, 2004).

Pilot Test

We administered the ALS to 87 Advanced Individual Training (AIT) students (E1-E3) and 44 NCOs serving as AIT instructors (E5-E7) during pilot testing. Preliminary data analysis showed good variability in responses and acceptable internal consistency reliability estimates for ALS scales. Moreover, correlations among the scales were not too large (i.e., most *r*s < .50), which was a concern because many of the constructs these scales are intended to measure are interrelated (e.g., satisfaction and commitment). Given the nature of the sample (i.e., modest sample of less experienced Soldiers who may not have known enough about the Army to make valid attitude ratings), we did not revise the ALS based on these results. We did, however, make some minor revisions based on feedback from respondents about the clarity of instructions, required reading level, and the items themselves.

Field Test

Sample

A total of 330 Soldiers completed the ALS during the criterion field test.[23] There were very little missing data, with 99% of the Soldiers responding to all 105 ALS items. We did, however, eliminate the responses of two Soldiers who test administrators flagged as having questionable ALS data. Thus, the analysis sample comprised 328 cases.

Analyses

For each ALS scale, we began by examining item-level statistics, including means, standard deviations (*SD*s), and item deletion statistics, to identify problematic items. We also used exploratory factor analysis (EFA) to assess the dimensionality of items comprising each scale. All EFA results reported in this chapter are based on a principal axis factor analysis with

[23] Information on the demographic characteristics of Soldiers who completed the measures discussed in this chapter is provided in Chapter 2.

oblique rotation of factors. We then computed descriptive statistics and intercorrelations for the revised scales. In addition, when multiple scales were hypothesized to comprise a higher-order construct (e.g., organizational commitment), we used confirmatory factor analysis (CFA) to assess the fit of the measurement model. These analyses were conducted with LISREL 8.3 (Jöreskog & Sörbom, 1996) on the covariance matrices using maximum likelihood estimation. Below we describe the analysis results for each set of ALS scales.

Results

Satisfaction with Army. The ALS measures six facets of satisfaction with the Army, including satisfaction with supervision (5 items), peers (4 items), work itself (7 items), promotions (4 items), pay and benefits (5 items), and the Army in general (10 items). Analyses revealed that the psychometric characteristics of these six scales were very good. In fact, no problematic items were found. Table 7.1 displays descriptive statistics, reliability estimates, and intercorrelations for these scales, along with the other scales (revised as necessary) that comprise the ALS. Internal consistency reliability estimates (alpha) for the satisfaction scales ranged from .82 (Satisfaction with Peers) to .93 (Satisfaction with Pay and Benefits). Scale correlations were quite modest, ranging from .20 between Peers and Pay and Benefits and .51 between Pay and Benefits and the Army in General.

CFA was used to assess the fit of a model comprising the six latent satisfaction factors. Overall, the fit statistics indicated an acceptable fit for the 6-factor model (see Table 7.2). For comparison, we also fitted a single-factor model (representing an overall satisfaction factor) by constraining the covariances among the six satisfaction factors to 1.0. Doing so allowed us to empirically compare the fit of the single satisfaction factor to that of the original 6-factor model (i.e., because the "constrained" model is nested within the original model). Analysis of the data revealed that including these constraints resulted in a substantial decrement in fit relative to the 6-factor model ($\Delta\chi^2$ (15) = 5,056.91, $p < .01$).

Organizational commitment. The ALS measures three facets of organizational commitment: affective (9 items), continuance (9 items), and normative commitment (7 items). Simply put, affective, continuance, and normative commitment, respectively, reflect the extent to which individuals want to remain with an organization (e.g., because they are emotionally attached to the organization), need to remain with an organization (e.g., because they lack alternatives), and feel obligated to remain with an organization (e.g., for the opportunities the organization has given them). Analysis of the data revealed six items across the three scales (i.e., 2 affective items, 1 continuance item, and 3 normative items) that had either low item-scale correlations and/or that correlated more highly with one of the other two commitment scales. Interestingly, the problematic items included all four reversed scored commitment items. In addition, all of these items were also somewhat problematic in the pilot test data. As a result, these six items were excluded from further analyses.

The psychometric characteristics of the revised commitment scales were acceptable (see Table 7.1), although the scale scores were rather highly intercorrelated. CFA was used to determine whether the data supported three distinct commitment factors. Results indicated a marginally acceptable fit for the 3-factor model (see Table 7.2). Some prior research (e.g., Ko,

Price, & Mueller, 1997) has found a lack of discriminant validity between affective and normative commitment. Given this, and the fact that these two scales were fairly highly related in the current sample ($r = .62$), we also assessed the fit of a 2-factor model in which the affective and normative commitment items comprised a single factor. In addition, we fitted a 1-factor model representing an overall commitment factor. Analyses revealed that the a priori 3-factor model fit significantly better than both the 2-factor model ($\Delta\chi^2$ (2) = 343.41, $p < .01$) and the single-factor model ($\Delta\chi^2$ (2) = 722.12, $p < .01$).

Perceived fit. The ALS assesses perceived fit with Soldiers' MOS (8 items) and the Army in general (9 items). Item analysis revealed five items (2 MOS fit and 3 Army fit) that, if deleted, would notably increase the internal consistency of their respective scales. Thus, these six items were excluded from subsequent analyses. The revised fit scales exhibited good measurement properties (see Table 7.1). CFA was used to assess the fit of the 2-factor model of MOS and Army fit. This model fit the data fairly well (see Table 7.2). For comparison, we created a single-factor model (by constraining the covariance between the two factors to 1.0), but found that doing so resulted in a significant decrement in fit relative to the 2-factor model ($\Delta\chi^2$ (1) = 323.00, $p < .01$).

Perceived stress. The Perceived Stress scale of the ALS comprises 10 items. Analyses revealed three items with relatively low item-scale correlations. EFA showed that two of these items, which dealt with relocating while in the Army, loaded a second, largely independent factor. Given this, these items were eliminated. The third item (which assessed the demands of the Army and family life) did not load strongly on either factor and also was eliminated. This left seven items, which produced a stress scale with acceptable psychometric characteristics (e.g., $\alpha = .76$).

Career intentions. The ALS includes three scales that measure different aspects of Army career intentions, including Attrition Cognitions (4 items), Re-Enlistment Intentions (3 items), and Army Career Intentions (3 items). One of the attrition cognitions items (likelihood of leaving Army if offered a good civilian job) did not contribute to scale reliability and thus was excluded from further analyses. One of the career intentions items (expected years of active duty) was excluded for the same reason.

The measurement characteristics of the final career intentions scales were quite good (see Table 7.1). However, the Re-enlistment and Army Career Intentions scales were highly correlated ($r = .79$), and therefore we were concerned these two scales might not be measuring distinct constructs. Thus, we used CFA to assess the fit of alternative models of the career intentions scales. We began by fitting a 3-factor model representing the three scales. This model fit the data quite well (see Table 7.2), despite the notable relationship ($r = .87$) between the re-enlistment and career factors. We then constrained the covariance between these two factors to 1.0 to create an overall "continuance intentions" factor. However, results revealed a significant decrement in fit relative to the original 3-factor model ($\Delta\chi^2$ (1) = 83.81, $p < .01$). Based on these results, we to retained all three scales.

Table 7.1. Descriptive Statistics, Reliability Estimates, and Intercorrelations for ALS Scale Scores

Scale	Items	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Satisfaction																			
1 Supervision	5	3.05	0.97	(.89)															
2 Peers	4	3.59	0.79	.39	(.82)														
3 Work Itself	7	2.83	0.95	.42	.38	(.91)													
4 Promotions	4	2.91	1.08	.37	.35	.40	(.89)												
5 Pay and Benefits	5	2.79	1.06	.27	.20	.35	.34	(.93)											
6 Army in General	10	2.86	0.74	.39	.31	.49	.39	.51	(.85)										
Organizational Commitment																			
7 Affective Commitment	7	2.77	0.83	.29	.33	.37	.36	.31	.63	(.86)									
8 Continuance Commitment	8	2.37	0.89	.08	.09	.22	.15	.19	.48	.56	(.87)								
9 Normative Commitment	4	1.88	0.90	.21	.10	.32	.23	.22	.50	.62	.66	(.85)							
Perceived Fit																			
10 MOS Fit	6	2.97	1.00	.16	.22	.50	.28	.22	.45	.40	.24	.32	(.88)						
11 Army Fit	6	3.05	0.84	.33	.31	.48	.32	.33	.67	.79	.52	.52	.59	(.80)					
12 Perceived Stress	7	3.31	0.69	-.32	-.18	-.34	-.26	-.31	-.54	-.45	-.18	-.32	-.45	-.56	(.76)				
Career Intentions																			
13 Attrition Cognitions	3	2.23	0.99	-.19	-.31	-.28	-.28	-.16	-.40	-.53	-.29	-.27	-.36	-.57	.34	(.69)			
14 Re-Enlistment Intentions*	3	1.90	0.97	.21	.19	.31	.22	.20	.51	.59	.60	.63	.35	.61	-.34	-.34	(.89)		
15 Army Career Intentions	2	1.83	1.12	.15	.10	.30	.19	.18	.45	.54	.51	.58	.39	.57	-.33	-.29	.79	(.90)	
16 Army Values	7	4.19	0.85	.17	.29	.24	.19	.17	.32	.45	.20	.17	.30	.45	-.34	-.41	.17	.17	(.91)

Note. $n = 328$. Items = number of items comprising each final scale. All items rated on a 5-point Likert-type scale. *Includes one item recoded to a 3-point scale. Internal consistency reliability estimates (alpha) are shown along the diagonal in parentheses. Correlations with an absolute value of .15 and greater are significant ($p < .05$).

Table 7.2. Fit Statistics for a Priori and Alternative Models of the ALS Scales

Scales/Model	χ^2	*df*	GFI	AGFI	CFI	RMSEA
Satisfaction with Army						
6 Factors	983.51	545	.85	.83	.93	.051
1 Factor	6063.47	560	.49	.42	.47	.173
Organizational Commitment						
3 Factors	427.44	149	.88	.85	.91	.076
2 Factors	770.85	151	.80	.75	.85	.112
1 Factor	1149.56	152	.73	.66	.78	.142
Perceived Fit						
2 Factors	188.03	53	.91	.87	.92	.088
1 Factor	509.53	54	.79	.70	.80	.161
Career Intentions						
3 Factors	47.46	17	.96	.93	.98	.074
2 Factors	131.27	18	.91	.82	.93	.139

Note. $n = 328$. For each set of scales, the fit statistics for the a priori model appear first, followed by the statistics for the alternate model(s). χ^2 = chi-square statistic. *df* = degrees of freedom. GFI = goodness-of-fit-index. AGFI = adjusted goodness-of-fit index. CFI = comparative fit index. RMSEA = root-mean-square error of approximation. All χ^2 are statistically significant ($p < .01$).

Army Values. The final seven ALS items comprise the Army Values scale. EFA of these data revealed a strong single factor ($\alpha = .91$) with no problematic items. As expected, ratings of the value items were generally quite high ($M = 4.19$). There was, however, a decent amount of variability in responses across the seven items (*SD*s = 0.96 to 1.15).

Scale intercorrelations. Given the similarity of constructs assessed by some of the ALS scales, we were concerned about overly high relations among scale scores. The scale correlations displayed in Table 7.1 suggest that this is not a significant issue. Intercorrelations ranged from .08 (Satisfaction with Supervision and Continuance Commitment) to .79 (Affective Commitment and Perceived Fit with Army).[24] The mean absolute correlation among the 16 ALS scales was .37. The mean correlation between scales designed to measure facets of the same construct (.44) was significantly larger ($p < .05$) than the mean correlation between scales designed to measure different constructs (.35).

We also performed an EFA of the 16 ALS scales. Results revealed that the data were best described by three factors. The first was a "continuance" factor that comprised four scales: Continuance Commitment, Normative Commitment, Re-Enlistment Intentions, and Army Career Intentions. The second was a "satisfaction" factor, which included the six satisfaction scales. The third factor comprised the remaining ALS scales, with Army Values, Attrition Cognitions, and Perceived Fit with Army having the strongest loadings on this factor.

Subgroup effect sizes. Finally, we examined subgroup differences on the ALS scales by gender, race, and MOS. The results of these analyses are presented in Tables 7.3, 7.4, and 7.5, respectively. In terms of gender, there were only three statistically significant mean differences,

[24] It is interesting that Affective Commitment and Perceived Fit with Army correlated higher with one another than with other facets within their respective constructs.

and the effect sizes associated with those differences were only moderately large. Specifically, females had higher scores on three of the satisfaction scales, namely Work Itself (d = 0.31), Promotions (d = 0.36), and Pay and Benefits (d = 0.45). Minimal differences were found with regard to race. There were no statistically significant mean differences between the scale scores of White and Black Soldiers, and there was only one significant difference between Whites and Hispanics whereby Hispanic Soldiers had higher scores than White Soldiers on MOS Fit (d = 0.32).

Table 7.3. ALS Scale Scores by Gender

		Male		Female	
Scale	d_{FM}	*M*	*SD*	*M*	*SD*
Satisfaction					
Supervision	0.13	3.04	0.95	3.16	1.01
Peers	0.12	3.58	0.78	3.68	0.83
Work Itself	**0.31**	2.79	0.96	3.09	0.88
Promotions	**0.36**	2.85	1.08	3.24	1.05
Pay and Benefits	**0.45**	2.71	1.04	3.18	1.07
Army in General	0.10	2.85	0.76	2.92	0.64
Organizational Commitment					
Affective Commitment	-0.14	2.79	0.84	2.67	0.79
Continuance Commitment	0.10	2.37	0.91	2.45	0.81
Normative Commitment	-0.17	1.91	0.93	1.75	0.71
Perceived Fit					
MOS Fit	-0.05	2.93	0.98	2.88	0.84
Army Fit	0.11	3.03	0.85	3.13	0.79
Perceived Stress	0.15	3.29	0.67	3.40	0.76
Career Intentions					
Attrition Cognitions	0.06	2.21	0.99	2.28	0.99
Re-Enlistment Intentions	0.13	1.86	0.98	1.98	0.95
Army Career Intentions	-0.02	1.84	1.14	1.82	1.03
Army Values	-0.04	4.19	0.86	4.15	0.75

Note. n_{Male} = 280, n_{Female} = 47. d_{FM} = Effect size for Female-Male mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., Males) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

In contrast, numerous statistically significant MOS differences emerged on the ALS. The main finding was that 31U Soldiers had more positive attitudes and intentions than Army-wide and 11B Soldiers. In fact, the 31U scores were significantly higher than the Army-wide and 11B scores on 9 and 11 of the 16 ALS scales, respectively. In addition, Army-wide scores were significantly higher than 11B scores on three of the ALS satisfaction scales (i.e., Supervision, Peers, and Pay and Benefits). The largest effect size differences were that 31U Soldiers were notably more satisfied with pay and benefits and supervision, respectively, than were 11B Soldiers (d = 0.80 and 0.76). In addition, 31U Continuance Commitment scale scores were much higher than both Army-wide and 11B scores on this scale (d = 0.70 and 0.61).

Table 7.4. ALS Scale Scores by Race/Ethnic Group

			White		Black		White Non-Hispanic		Hispanic	
Scale	d_{BW}	d_{HW}	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Satisfaction										
Supervision	0.18	0.29	2.95	0.97	3.13	0.95	2.97	0.94	3.24	1.04
Peers	0.21	0.15	3.55	0.85	3.73	0.69	3.56	0.84	3.69	0.82
Work Itself	0.26	0.28	2.71	0.96	2.96	0.90	2.70	0.95	2.97	0.97
Promotions	0.24	0.19	2.82	1.12	3.09	1.00	2.82	1.11	3.03	1.08
Pay and Benefits	0.27	-0.01	2.72	1.06	3.00	1.16	2.73	1.05	2.73	1.12
Army in General	-0.08	0.19	2.84	0.74	2.79	0.68	2.84	0.73	2.98	0.78
Organizational Commitment										
Affective Commitment	-0.30	-0.04	2.81	0.84	2.56	0.86	2.81	0.85	2.78	0.74
Continuance Commitment	0.00	0.10	2.36	0.91	2.37	0.85	2.36	0.91	2.46	0.87
Normative Commitment	-0.13	0.09	1.85	0.90	1.73	0.67	1.85	0.90	1.93	0.86
Perceived Fit										
MOS Fit	-0.04	**0.32**	2.89	1.00	2.86	0.85	2.87	1.00	3.19	0.92
Army Fit	-0.15	0.16	3.05	0.85	2.92	0.92	3.04	0.85	3.18	0.76
Perceived Stress	0.12	-0.26	3.34	0.72	3.42	0.62	3.35	0.71	3.16	0.65
Career Intentions										
Attrition Cognitions	0.11	0.03	2.16	1.02	2.28	0.92	2.18	1.01	2.21	1.06
Re-Enlistment Intentions	-0.02	-0.01	1.88	0.99	1.86	0.93	1.86	0.97	1.85	1.03
Army Career Intentions	-0.03	-0.07	1.86	1.16	1.82	1.08	1.84	1.12	1.76	1.14
Army Values	-0.09	0.09	4.20	0.85	4.13	0.75	4.19	0.82	4.26	0.93

Note. $n_{White} = 209$. $n_{Black} = 53$. $n_{White\ Non\text{-}Hispanic} = 197$. $n_{Hispanic} = 50$. d_{BW} = Effect size for Black-White mean difference. d_{HW} = Effect size for Hispanic-White Non-Hispanic mean difference. Effect sizes calculated as (mean of non-referent group - mean of referent group)/*SD* of referent group. Referent groups (e.g., White) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

Table 7.5. ALS Scale Scores by MOS

	$d_{AW\text{-}11B}$	$d_{AW\text{-}31U}$	$d_{31U\text{-}11B}$	AW *M*	AW *SD*	11B *M*	11B *SD*	31U *M*	31U *SD*
Satisfaction									
Supervision	**0.27**	**-0.50**	**0.76**	3.09	1.01	2.83	0.87	3.57	0.83
Peers	**0.27**	-0.33	**0.60**	3.66	0.75	3.45	0.81	3.92	0.63
Work Itself	0.03	-0.16	0.19	2.79	0.96	2.76	0.90	2.95	1.03
Promotions	-0.14	**-0.52**	0.38	2.61	1.03	2.76	1.09	3.17	0.90
Pay and Benefits	**0.31**	**-0.49**	**0.80**	2.77	1.07	2.44	1.03	3.29	1.04
Army in General	0.20	-0.22	**0.42**	2.93	0.81	2.78	0.76	3.09	0.66
Organizational Commitment									
Affective Commitment	-0.01	**-0.53**	**0.52**	2.72	0.83	2.73	0.84	3.16	0.71
Continuance Commitment	-0.09	**-0.70**	**0.61**	2.33	0.89	2.41	0.95	2.96	0.88
Normative Commitment	-0.18	**-0.61**	**0.43**	1.79	0.80	1.95	1.00	2.34	0.93
Perceived Fit									
MOS Fit	-0.16	-0.41	0.25	2.72	0.95	2.88	0.96	3.12	0.82
Army Fit	0.11	**-0.50**	**0.61**	3.02	0.83	2.92	0.90	3.44	0.68
Perceived Stress	-0.09	0.34	**-0.44**	3.33	0.67	3.39	0.68	3.09	0.64
Career Intentions									
Attrition Cognitions	-0.11	0.12	-0.24	2.20	0.88	2.31	1.12	2.08	0.90
Re-Enlistment Intentions	-0.02	**-0.50**	**0.49**	1.86	0.99	1.88	1.00	2.35	1.10
Army Career Intentions	-0.06	**0.46**	0.40	1.79	1.15	1.86	1.18	2.31	1.17
Army Values	0.18	-0.33	**0.50**	4.19	0.83	4.04	0.96	4.46	0.57

Note. $n_{AW} = 112$. $n_{11B} = 122$. $n_{31U} = 29$. AW = Army-Wide. 11B = Infantryman. 31U = Signal Support Systems Specialist. $d_{AW\text{-}11B}$ = Effect size for AW-11B mean difference. $d_{AW\text{-}31U}$ = Effect size for AW-31U mean difference. $d_{31U\text{-}11B}$ = Effect size for 31U-11B mean difference. Effect sizes calculated as (mean of non-referent group - mean of referent group)/*SD* across all Soldiers. Referent groups (e.g., 11B) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

Discussion

In summary, the overall findings from field test of the ALS were very promising. For example, all of the ALS scales produced sufficient variance in scores and had acceptable estimates of internal consistency reliability. We also found evidence for the intended factor structure of the ALS scales, and with few exceptions, correlations among scales were rather modest. In addition, there were very few notable scale score differences with regard to gender and race. As such, we recommend retaining all 16 ALS scales for the concurrent validation. We will, however, revise or eliminate a handful of items that did not perform well in the criterion field test data collection. We anticipate the revised ALS to include about 95 items and to take 20-25 minutes to complete.

Despite our optimism, we do have a few minor concerns with the ALS. The organizational commitment measures were perhaps the most problematic (e.g., in terms of items cross-loading on multiple factors) ALS scales within both the pilot and field test data collections. However, we think that rewording a few of the items and eliminating a couple of others will improve the measurement of the commitment facets. A related concern is the high correlation

between the Affective Commitment and Army Fit scale scores. Indeed, the correlation (corrected for unreliability in both measures) between these scale scores was .95. Nonetheless, the constructs we intend to assess with these two scales are sufficiently different, and there remains enough unique variance to justify their administration in the concurrent validation. A final concern with the ALS is some of the large MOS differences observed in the field test data, namely the consistently higher scores for 31U Soldiers compared to Army-wide and 11B Soldiers. Although such mean differences might not affect correlations between the P-E fit predictors and criteria, they could (e.g., they could produce intercept biases for variables used to predict attitudinal criteria). Thus, we will be particularly cognizant of MOS differences (and their effects) during the concurrent validation.

Future Army Life Survey

Description of Measure

We also developed a future-oriented measure that assesses how Soldiers' attitudes and perceptions of work might change as the Army transforms to the Future Force. These future-state attitudes/intentions are assessed in the Future Army Life Survey (FALS). The 29-item FALS was designed to assess many of the same attitudinal and intention constructs measured in the ALS. Specifically, the FALS measures (a) expected satisfaction with future conditions, (b) perceived stressfulness of future conditions, (c) expected performance under future conditions, (d) perceived fit with the future Army (i.e., abilities-demands and needs-supplies fit), and (e) re-enlistment and career intentions given future conditions. To give Soldiers a context for responding to the FALS, they are asked to read descriptions of anticipated future Army conditions (e.g., frequent change, continuous learning) prior to completing the survey. These conditions were based on the Select21 future-oriented job analysis (Sager, Russell, R.C. Campbell, & Ford, 2005).

Unlike the ALS, an issue we confronted in developing the FALS was how to assess Soldiers' attitudes about future conditions they have yet to experience. The descriptions of anticipated future Army conditions limit the constructs we could reasonably assess in the FALS. Specifically, without broader descriptions of the future Army work environment (which are not available), the content of the FALS was limited to constructs that deal with work demands in general (e.g., expected satisfaction with and stressfulness of the work described). Using the FALS to assess constructs such as affective commitment and satisfaction with supervision and pay (which are likely to be influenced by broader characteristics of the future environment) based on scenarios that focus on the shifting demands of work would be problematic. As such, several constructs assessed in the ALS were excluded from the FALS.

Field Test Results

Sample

Data on the FALS were gathered from 326 Soldiers. Criterion field test problem logs revealed no issues with these Soldiers' FALS data.

Scale Development

Given the FALS was a new measure, we used EFA to examine the factor structure underlying its 29 items. First, we conducted an EFA using the principal axis factor extraction method with oblique rotation. In this initial analysis, we did not specify the number of factors to be extracted. Examination of these results indicated that six factors had eigenvalues over 1.0, yet the scree plot revealed a distinct drop off in the magnitude of eigenvalues after the third factor. In light of these results, we conducted three follow-up EFA that constrained the solution to two, three, and four factors, respectively. Comparing the results from these analyses, we concluded that the 3-factor solution provided the most interpretable factors and resulted in few cross loadings. The pattern matrix resulting from the 3-factor model and the associated eigenvalues and communalities are shown in Table 7.6.

The factors underlying the FALS data appear to reflect three constructs: (a) overall fit with the future Army (Future Fit), (b) perceived stressfulness of the future Army (Future Stress), and (c) future continuance intentions (Future Continuance). Examination of the eigenvalues indicated the first factor accounted for most of shared variance among the items. The presence of a strong first factor among the items is not surprising in that perceptions of overall fit likely influence Soldiers' perceptions of future stress and continuance.

On this basis, we formed three FALS scales by taking the average of items that had loadings of .30 or greater on their respective factor. Items with loadings of .30 or greater on multiple factors were omitted. Internal consistency reliability estimates and item-deleted alphas for each of these scales were evaluated. Based on these results, we retained 13 items for the Future Fit scale (α = .90), six items for the Future Stress scale (α = .78), and seven items for the Future Continuance scale (α = .90). Loadings for items included in the final scales are bolded in Table 7.6. As a follow-up to forming these scales, we used CFA to assess the fit of a 3-factor model to the retained FALS items. Overall, the fit statistics indicated fair levels of fit (e.g., CFI = .83, RMSEA = .08). Furthermore, factor score determinacies were all close to 1.0 (Future Fit = .96, Future Stress = .90, Future Continuance = .96), suggesting the FALS items provide good assessments of the three factors.[25]

Next, we generated descriptives statistics, intercorrelations, and subgroup means for the FALS scales (see Tables 7.7 through 7.10). All three FALS scales exhibited good variability. Examination of their score distributions and skewness statistics indicated that Future Stress and Future Continuance were normally distributed, and Future Fit exhibited a slight negative skew (*Skew* = -0.39). The scales were moderately to highly correlated, with Future Fit and Future Continuance being the most correlated (r = .66). Although sizeable, this correlation is not so high as to suggest that the scales fail to offer unique variance that could lead to different correlations with various predictor variables. For example, although Future Fit and Continuance share about 45% of their variance, the remaining 65% of their variance is unique.

[25] Factor score determinacies index how well a factor is measured by the items that load on it. They reflect the correlation between the estimated and true factor scores, or the square root of factor "reliability" (Muthen & Muthen, 2001).

Table 7.6. Pattern Matrix, Eigenvalues, and Communalities for FALS 3-Factor Solution

	Factor			
Item	1	2	3	h^2
Dealing with Info: I would perform well under this condition.	**0.79**	-0.02	0.07	0.57
Continuous Learning: I would perform well under this condition.	**0.69**	0.03	-0.09	0.56
I am confident that I could succeed under these conditions.	**0.68**	0.00	-0.04	0.50
Working Independently: I would perform well under this condition.	**0.65**	-0.01	0.01	0.42
I have the skills and abilities to succeed under these conditions.	**0.65**	-0.08	0.08	0.41
I would perform well under these conditions.	**0.63**	0.05	-0.15	0.51
Dealing with Info: How satisfied would you be working under this condition?	**0.63**	0.01	-0.08	0.45
It would be difficult for me to adapt to the demands of the future Army.	**-0.62**	0.14	-0.11	0.39
Continuous Learning: How satisfied would you be working under this condition?	**0.61**	0.04	-0.13	0.47
Frequent Change: I would perform well under this condition.	**0.52**	-0.08	-0.11	0.40
Working Independently: How satisfied would you be working under this condition?	**0.49**	-0.02	-0.03	0.26
I would not be a good fit for the future Army.	**-0.46**	-0.05	0.20	0.34
If current conditions were like these, I would not have joined the Army.	**-0.31**	0.07	0.24	0.27
Frequent Change: How satisfied would you be working under this condition?	0.28	-0.14	-0.27	0.31
Dealing with Info: How stressful would you find it to work under this condition?	-0.06	**0.67**	0.00	0.49
Continuous Learning: How stressful would you find it to work under this condition?	-0.07	**0.66**	0.04	0.50
Compared to the current Army, how stressful would you find the future Army?	0.17	**0.60**	0.16	0.36
Working Independently: How stressful would you find it to work under this condition?	-0.22	**0.60**	-0.21	0.42
Frequent Change: How stressful would you find it to work under this condition?	-0.05	**0.57**	-0.03	0.35
The future Army sounds like a stressful place to work.	-0.01	**0.42**	0.24	0.29
I would feel good about making the Army a career under these conditions.	-0.03	-0.03	**-0.81**	0.64
I would feel good about re-enlisting if current conditions were like these.	0.02	-0.05	**-0.77**	0.64
The future Army sounds like it would meet my needs and interests.	0.03	-0.05	**-0.71**	0.56
The future Army sounds like a good place to work.	0.14	0.02	**-0.69**	0.62
Compared to the current Army, would you be more or less likely to re-enlist in the future Army?	0.01	-0.16	**-0.68**	0.56
Compared to the current Army, would you be more or less likely to make the future Army a career?	0.12	-0.04	**-0.64**	0.54
The future Army sounds like an interesting place to work.	0.23	0.08	**-0.52**	0.44
I would be satisfied working under these conditions.	0.41	0.04	-0.44	0.55
Compared to the current Army, how satisfied would you be in the future Army?	0.32	-0.01	-0.42	0.45
Final Eigenvalues	10.19	1.78	1.33	

Note. h^2 = Communalities (proportion of variance accounted for in item by the three-factor solution. Loadings for items included in FALS scales corresponding to each factor are bolded.

With regard to subgroup differences, males had significantly higher Future Fit scores than females (d = -0.33), and White Soldiers had significantly higher Future fit scores than Black Soldiers (d = -0.43). Though these differences were statistically significant, the magnitudes of these effects sizes are relatively small (Cohen, 1992). Only one statistically significant difference was found with regard to MOS–31Us had significantly higher Future Continuance scores than Soldiers in Army-wide MOS (d = -0.46). No other effect sizes for MOS comparisons exceeded 0.40.

Table 7.7. Descriptive Statistics, Reliability Estimates, and Intercorrelations for FALS Scale Scores

Scale	*M*	*SD*	1	2	3
1 Future Fit	3.55	0.64	(.90)		
2 Future Stress	3.15	0.67	-.42	(.78)	
3 Future Continuance	3.02	0.86	.66	-.35	(.90)

Note. n = 326. Internal consistency reliability estimates (alpha) are shown along the diagonal in parentheses. All correlations are statistically significant, $p < .05$ (one-tailed).

Table 7.8. FALS Scale Scores by Gender

		Male		Female	
Scale	d_{FM}	*M*	*SD*	*M*	*SD*
Future Fit	**-0.33**	3.58	0.63	3.37	0.71
Future Stress	0.30	3.12	0.66	3.32	0.70
Future Continuance	-0.21	3.05	0.87	2.87	0.82

Note. n_{Male} = 278, n_{Female} = 47. d_{FM} = Effect size for Female-Male mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., Males) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

Table 7.9. FALS Scale Scores by Race/Ethnic Group

			White		Black		White Non-Hispanic		Hispanic	
Scale	d_{BW}	d_{HW}	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Future Fit	**-0.43**	0.03	3.61	0.61	3.35	0.80	3.58	0.65	3.60	0.65
Future Stress	0.16	-0.13	3.15	0.65	3.26	0.78	3.18	0.66	3.09	0.66
Future Continuance	-0.12	0.05	3.07	0.82	2.98	0.99	3.03	0.81	3.07	1.04

Note. n_{White} = 207. n_{Black} = 52. $n_{White\ Non\text{-}Hispanic}$ = 195. $n_{Hispanic}$ = 52. d_{BW} = Effect size for Black-White mean difference. d_{HW} = Effect size for Hispanic-White Non-Hispanic mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., White) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

Table 7.10. FALS Scale Scores by MOS

				AW		11B		31U	
Scale	$d_{AW\text{-}11B}$	$d_{AW\text{-}31U}$	$d_{31U\text{-}11B}$	*M*	*SD*	*M*	*SD*	*M*	*SD*
Future Fit	0.07	-0.18	0.24	3.54	0.69	3.49	0.62	3.65	0.52
Future Stress	0.03	0.31	-0.28	3.18	0.69	3.15	0.63	2.97	0.64
Future Continuance	-0.16	**-0.46**	0.30	2.94	0.88	3.08	0.85	3.34	0.81

Note. n_{AW} = 113. n_{11B} = 119. n_{31U} = 29. AW = Army-Wide. 11B = Infantryman. 31U = Signal Support Systems Specialist. $d_{AW\text{-}11B}$ = Effect size for AW-11B mean difference. $d_{AW\text{-}31U}$ = Effect size for AW-31U mean difference. $d_{31U\text{-}11B}$ = Effect size for 31U-11B mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* across all Soldiers. Referent groups (e.g., 11B) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

Correlations with ALS Scale Scores

In creating the FALS, one concern was whether its scales would be distinct from the ALS scales. Given that the descriptions of the future Army provided to Soldiers completing the FALS were necessarily partial descriptions of anticipated future Army conditions, much of the variance in FALS scales could simply reflects Soldiers' current attitudes towards the Army. To assess this possibility, we examined correlations between ALS and FALS scales (see Table 7.11).

Table 7.11. Correlations between ALS and FALS Scale Scores

	FALS Scale		
ALS Scale	Future Fit	Future Stress	Future Continuance
Satisfaction			
Supervision	.05	.06	.03
Peers	**.21**	-.06	**.18**
Work Itself	**.16**	-.09	**.23**
Promotions	**.21**	-.09	**.21**
Pay and Benefits	.08	**-.12**	**.16**
Army in General	**.29**	-.08	**.38**
Organizational Commitment			
Affective Commitment	**.41**	**-.17**	**.47**
Continuance Commitment	**.16**	-.01	**.38**
Normative Commitment	**.21**	**-.13**	**.45**
Perceived Fit			
MOS Fit	**.27**	-.09	**.30**
Army Fit	**.44**	**-.19**	**.45**
Perceived Stress	**-.31**	**.26**	**-.28**
Career Intentions			
Attrition Cognitions	**-.38**	**.14**	**-.28**
Re-Enlistment Intentions	**.28**	-.10	**.43**
Army Values	**.34**	**-.14**	**.26**

Note. $n = 324$. Bolded correlations are statistically significant, $p < .05$ (two-tailed).

Relations between ALS and FALS scales were small to moderate, with no correlation exceeding .47 in magnitude.[26] The strongest ALS correlates of Future Fit were current Perceived Army Fit ($r = .45$) and Affective Commitment ($r = .41$). Of the three FALS scales, Future Stress showed the least relationship to the ALS scales. The strongest ALS correlate of Future Stress was current Perceived Stress ($r = .26$). Lastly, several ALS scales exhibited notable correlations with Future Continuance. The strongest ALS correlates of Future Continuance were current Affective Commitment ($r = .47$), Normative Commitment ($r = .45$), Perceived Army Fit ($r = .45$), and Re-Enlistment Intentions ($r = .43$).

At the bivariate level, the FALS scales appear to be related to, yet distinct from, Soldiers' attitudes towards the current Army. Nevertheless, it is possible that combinations of the ALS scales could account for a large portion of the variance in the FALS scales. Such findings would suggest

[26] Given the relatively high reliability estimates for most of the ALS and FALS scales, correcting these observed correlations for unreliability did not notably increase relations.

that Soldiers' perceptions of the future Army are simply a function of their attitudes towards the current Army. To examine this possibility, we conducted three stepwise regression analyses using backward elimination. For each analysis, one of the FALS scales served as the criterion, and all of the ALS scales were initially entered as predictors. Only ALS scales with statistically significant ($p < .05$) beta weights were included in the final model for each FALS scale.

Results revealed that 26.5% of the variance in Future Fit ($R = .52$) was accounted for by six ALS scales (Perceived Fit with the Army, $\beta = .28$; Affective Commitment, $\beta = .18$; Continuance Commitment, $\beta = -.14$; Attrition Cognitions, $\beta = -.14$, Army Values, $\beta = .13$; Satisfaction with Supervision, $\beta = -.13$). Additionally, 9.1% of the variance in Future Stress ($R = .30$) was accounted for by two ALS scales (Perceived Stress, $\beta = .31$; Satisfaction with Supervision, $\beta = .16$). Lastly, 30.1% of the variance in Future Continuance ($R = .55$) was accounted for by four ALS scales (Perceived Fit with the Army, $\beta = .32$; Normative Commitment, $\beta = .31$; Satisfaction with Supervision, $\beta = -.19$; Satisfaction with Peers, $\beta = .13$).

Thus, although the proportions of variance accounted for in the FALS scales by the ALS scales were notable (particularly for Future Fit and Future Continuance), the vast majority of variance in the FALS scales remained unaccounted for by the ALS scales. Taken together, these findings suggest that Soldiers' perceptions of the future Army are more than simply a function of their attitudes towards the current Army.

Additional Analyses

In addition to developing and evaluating scales for potential use in the concurrent validation, another purpose behind examining the FALS data was to assess Soldiers' views with regard to anticipated future Army conditions, particularly relative to current conditions. Although much work has been done with regard to what the Army will be like in the future, little is known about how individual Soldiers feel about the anticipated changes. Four questions on the FALS asked Soldiers to indicate how they felt about anticipated future Army conditions relative to their feelings towards the current Army. Response distributions for these items are presented in Table 7.12.

Table 7.12. Response Distributions for Future vs. Current Army FALS Items

Item	Much More	Somewhat More	About the Same	Somewhat Less	Much Less
Compared to the current Army, how satisfied would you be in the future Army?	12.3	32.2	40.5	8.9	6.1
Compared to the current Army, how stressful would you find the future Army?	8.0	34.0	42.9	11.0	4.0
Compared to the current Army, would you be more or less likely to re-enlist in the future Army?	10.1	16.6	49.7	11.3	12.3
Compared to the current Army, would you be more or less likely to make the future Army a career?	8.0	15.0	52.5	10.7	13.8

Note. Cell values indicate the percentage of Soldiers who endorsed the given response option for each item.

Results showed that 42% of Soldiers indicated that they would find the future Army more stressful than the current Army, whereas only 15% of Soldiers indicated they would find the future Army less stressful. Despite this view of the future as more stressful, 45% of Soldiers indicated they would find the future Army more satisfying than the current Army, whereas only 15% of Soldiers indicated they would find the future Army less satisfying. With regard to re-enlistment and career intentions, roughly equal proportions of Soldiers indicated they would be more or less likely to re-enlist or pursue an Army career in the future. Caution should be taken in overinterpreting these results because they are based on single items, and they do not reveal Soldiers feelings towards the current Army. Nevertheless, they do provide a glimpse of how Soldiers might feel about anticipated future conditions relative to current conditions.

Discussion

In developing the FALS, a concern we had was that Soldiers' responses might primarily reflect their attitudes towards the current Army (assessed by the ALS). The results presented in this chapter suggest that the FALS not only appears distinct from the ALS, but also has other desirable properties. Specifically, factor analyses of the FALS items revealed a relatively clean three-factor solution, and scales based on that solution had relatively normal distributions and exhibited good levels of internal consistency reliability and variability. Given the future-oriented focus of Select21, as well as the relatively short time it takes to administer the FALS, we recommend administering this measure in its entirety during the concurrent validation. We expect the revised instrument to take 10-15 minutes to administer.

CHAPTER 8: DEVELOPMENT OF THE RATIONAL BIODATA INVENTORY (RBI)

Robert N. Kilcullen
U.S. Army Research Institute

Dan J. Putka, Rodney A. McCloy, and Chad H. Van Iddekinge
HumRRO

Background

Biodata tests are self-report multiple-choice questionnaires that attempt to measure the test-taker's prior behavior, experiences, and reactions to life events. Meta-analyses of the selection literature show that biodata effectively predict a wide variety of performance criteria (e.g., ratings of overall performance, advancement potential, commendations, sales volume, bonuses), with typical estimated validities in the .30s to .40s (Hunter & Hunter, 1984; Reilly & Chao, 1982; Schmitt, Gooding, Noe, & Kirsch, 1984). In addition to being useful as an initial selection screen, biodata instruments achieve similar validity estimates for predicting various criteria of supervisory and managerial performance (Owens, 1976; Reilly & Chao, 1982).

Historically, biodata instruments are empirically keyed or scored. Several mathematical methods of empirical keying have been derived, but each involves scoring item responses based on the strength of their statistical relation to a criterion. Typically, biodata items and the criterion measure of interest are administered to a large sample of subjects. Next, items on which the best and worst subjects differentially respond are identified and retained. Points are assigned to each item response based on the quality of applicants endorsing the response. The validity of the scoring key is then cross-validated on another similar sample of subjects.

Unfortunately, empirical keying strategies have serious drawbacks. They often show high validity initially, but suffer substantial shrinkage across samples and over time (Schwab & Oliver, 1974; Walker, 1985; White & Kilcullen, 1992). In addition, item selection and scoring are atheoretical, which makes it difficult to understand what is being measured or why different criterion groups respond differently to the biodata items (Mumford & Stokes, 1991).

Awareness of these problems led to increasing interest in rational keying strategies. This typically involves identifying constructs likely to predict the criterion of interest and writing biographical items to measure those predictor constructs (e.g., Emotional Stability, Conscientiousness). Item response weights are rationally assigned based upon the expected relations between the responses and the underlying construct. The scored item responses are then summed to form scale scores having substantive meaning. These scales typically show good convergent and discriminant validity, with personality "marker" scales measuring the same attributes and lower correlations with scales designed to detect socially desirable responding compared to the personality measures (Kilcullen, White, Mumford, & Mack, 1995).

The potential advantages of rational keying include a greater theoretical understanding of the phenomenon under study (Mumford & Stokes, 1991; Mumford, Uhlman, & Kilcullen, 1992). Additionally, rational keys typically yield criterion-related validity estimates that are comparable to those achieved with cross-validated empirical keys (Schoenfeldt, 1989; Uhlman, Reiter-Palmon, & Connelly, 1990) and tend to produce more stable validity estimates over time

(Clifton, Kilcullen, Reiter-Palmon, & Mumford, 1992; White & Kilcullen, 1992). In fact, in White and Kilcullen's research, item-level analyses revealed that stability in the empirical key was based on the degree to which it resembled the rational key. For these reasons, the rational keying approach was chosen as the method for developing and scoring the Select21 biodata test.

Constructs Targeted for Measurement

Rational scale construction requires identification of the psychological constructs to measure. In this regard, one consideration was to measure temperament constructs identified as important to future first-term Soldier performance in the Select21 job analysis (Sager, Russell, R.C. Campbell, & Ford, 2005). Another consideration was to leverage previously validated constructs and scales from two rational biodata tests – the Assessment of Right Conduct (ARC) and the Test of Adaptable Personality (TAP). ARI researchers designed the ARC to measure motivational attributes that lead to delinquent behavior. The ARC scales have been empirically linked to counterproductive behavior in a variety of Army settings (Kilcullen, White, Sanders, & Hazlett, 2003). Conversely, the TAP is oriented towards predicting job performance. It measures temperament constructs that predict the field performance of Special Forces Soldiers (Kilcullen, Goodwin, Chen, Wisecarver, & Sanders, 2002; Kilcullen, Mael, Goodwin, & Zazanis, 1999). Additional biodata scales validated for predicting job performance in leadership positions (Kilcullen, White, Zaccaro, & Parker, 2000) were also reviewed for possible inclusion in the Rational Biodata Inventory (RBI). The linkages between the initial set of constructs measured by the RBI and the Select21 KSAs are shown in Table 8.1[27].

RBI Scale	Corresponding Select21 KSA
Peer Leadership	Affiliation; Potency
Cognitive Flexibility	Intellectance
Achievement Orientation	Achievement Motivation
Fitness Motivation	Potency
Interpersonal Skills – Diplomacy	Affiliation
Stress Tolerance	Emotional Stability
Hostility to Authority	Dependability; Agreeableness
Self-Esteem	Self-Reliance; Emotional Stability
Narcissism	Dependability
Cultural Tolerance	Cultural Tolerance; Agreeableness
Internal Locus of Control	Locus of Control; Emotional Stability

Figure 8.1. Linkages between RBI scales and Select21 KSAs.

Table 8.1 lists the initial set of temperament constructs targeted for measurement by the RBI by source of content (e.g., TAP, ARC). It includes six scales from the TAP, two from the ARC, and three new scales targeting Select21 constructs not measured by previous biodata instruments.

[27] Two other scales were subsequently added to the RBI and retained on the final version. "Respect for Authority" corresponds to the Select21 KSA of "Dependability. "Army Identification" does not have a corresponding KSA.

Table 8.1. RBI Pilot Test Descriptive Statistics and Reliability Estimates

Scale	# Items	*M*	*SD*	*Alpha*
TAP				
1. Peer Leadership	11	3.33	0.57	0.78
2. Cognitive Flexibility	9	3.34	0.49	0.60
3. Achievement Orientation	9	3.69	0.57	0.76
4. Fitness Motivation	10	3.32	0.65	0.79
5. Interpersonal Skills – Diplomacy	5	3.58	0.79	0.74
6. Stress Tolerance	10	3.03	0.53	0.64
ARC				
7. Hostility to Authority	9	2.45	0.64	0.75
8. Self-Esteem	5	4.10	0.61	0.77
NEW				
9. Narcissism	9	3.45	0.50	0.61
10. Cultural Tolerance	6	3.67	0.71	0.71
11. Internal Locus of Control	8	3.48	0.59	0.69
OTHER				
12. Lie Scale	7	0.07	0.11	---

Note. n = 319.

Also included in the RBI is the Lie scale used in the TAP and the ARC to detect deliberate response distortion. Item scoring for the Lie scale is based on the endorsement of unlikely virtues. Previous research indicates that this scale shows good convergent and discriminant validity with a previously validated temperament scale measuring the same type of response distortion (Kilcullen et al., 1995). Since the goal of Select21 is to develop selection tests for operational use where faking on self-report measures is a concern, the Lie scale in this research was used as one criterion for eliminating pilot items. The goal is to use items that are not highly correlated with the Lie scale in the final version of the RBI with the idea that such items would be prone to have their responses distorted in an operational setting

Development of the Initial Item Pool

To construct the new biodata scales, a panel of psychologists reviewed the definitions of the constructs to be targeted by the new scales. Then independently, each psychologist generated several items referring to past behaviors and life events thought to be indicative of the construct in question. As a group, the panel then reviewed each new item for construct relevance, response variability, relevance to the intended population, readability, non-intrusiveness, and neutrality with respect to social desirability. A consensus decision was reached concerning best items for each construct, and the item response options were scored on a continuum to reflect the anticipated relationship between the responses and the construct.

The same procedure was followed previously for the TAP, ARC, and leadership scales. However, because these items had originally been created for different populations (e.g., Special Forces Soldiers), the panel of psychologists reviewed the items and reworded them to make them

more relevant to the population targeted with Select21. For example, the following rewording was made to an ARC Hostility to Authority item:

Old wording: At work, how often did your bosses enjoy giving people a hard time?

New wording: At school, how often did your teachers enjoy giving people a hard time?

Other previously developed biodata items were deleted either for being irrelevant to many new recruits (e.g., At work I take long lunches or extended breaks now and then) or to make the RBI conform to testing time requirements. The result was a 137-item biodata test.

Pilot Testing

Sample

Data on the RBI were gathered from 332 new recruits at Forts Knox and Jackson. Thirteen cases were discarded because they did not respond seriously to the test questions. In some cases, this was observed directly by test administrators. In other cases, it was detected based on the subject's pattern of responses (e.g., choosing the same response to every item) to the RBI. The result was an RBI analysis sample of 319 cases.

Analyses and Results

Psychometric analyses were performed so that the scales could be refined. The internal consistency reliability of each scale was assessed, and items that did not correlate strongly with their scale were discarded. In addition, five scales were dropped due to either poor internal consistency or a high correlation (i.e., $r > .60$) with another RBI scale. In every case, scales that correlated highly came from different biodata tests. Generally speaking, preference was given to retaining the TAP scales in instances when they correlated highly with other scales because the TAP has the strongest track record for predicting Soldier job performance. The resulting RBI comprised 98 items. Descriptive statistics for the resulting RBI scales are presented in Table 8.1. All of the scales yielded internal consistency reliability estimates in excess of .60 with a median of .74.

To assess convergent validity of the RBI scales, we correlated the RBI scale scores with scale scores from the International Personality Item Pool's (IPIP) 50-item "Big Five personality factor" marker test (International Personality Item Pool, 2001). The "IPIP" (as we subsequently refer to it) was administered to recruits as part of the pilot test to provide a marker measure for use in the development of the Select21 temperament measures (i.e., the RBI and Work Suitability Inventory [discussed in Chapter 9]). Each Big Five factor on the IPIP is assessed using 10 items (five positively keyed items and five negatively keyed items). The items on the IPIP are simple phrases that describe a person's behavior (e.g., pay attention to details.). Recruits were asked to rate the degree to which each statement provided an accurate description of themselves using a 5-point scale ranging from *Very Inaccurate* (1) to *Very Accurate* (5). Past research has indicated that the IPIP provides a reliable and construct-valid assessment of the Big Five (IPIP, 2001). Psychometric statistics and scale correlations for the IPIP are shown in Table 8.2.

Table 8.2. IPIP Descriptive Statistics and Intercorrelations

Scale	*M*	*SD*	1	2	3	4	5
1. Extraversion	3.25	0.83	---				
2. Agreeableness	3.66	0.66	**.42**	---			
3. Conscientiousness	3.70	0.65	.08	**.25**	---		
4. Emotional Stability	3.31	0.75	**.21**	**.23**	**.39**	---	
5. Intellectance	3.58	0.60	**.27**	**.37**	**.25**	**.23**	---

Note. $n = 329$. Statistically significant correlations are bolded, $p < .05$ (two-tailed).

The relationship between the RBI scales and the International Personality Item Pool (2001) Big Five marker scales are illustrated in Table 8.3. The pattern of relations generally is consistent with our expectations about what constructs are measured by the RBI scales. Reading across the rows of Table 8.3, the underlined correlation indicates the hypothesized relationship between the RBI scale and the Big Five construct. For example, the third row of Table 8.3 indicates that RBI Achievement Orientation was expected to correlate the highest with International Personality Item Pool (IPIP) Conscientiousness, which it did. As well, the RBI Peer Leadership and Interpersonal Skills – Diplomacy scales correlated the strongest with IPIP Extraversion, as expected. Likewise, RBI Cognitive Flexibility correlated the highest with IPIP Openness, and RBI Cultural Tolerance correlated the strongest with IPIP Agreeableness. In addition, RBI Stress Tolerance and Internal Locus of Control correlated the strongest with IPIP Emotional Stability. Although not all of the correlations were as expected – for example, the correlation between RBI Self-Esteem and IPIP Conscientiousness was as strong as the anticipated relationship between RBI Self-Esteem and IPIP Emotional Stability – empirical support was obtained for seven of the nine RBI-IPIP relationships that were expected to be the strongest.

Table 8.3. Correlations of RBI Scales with IPIP Big Five Marker Scales

	IPIP Scales				
RBI Scales	Extr	Agr	Cons	ES	Intel
1. Peer Leadership	**.48**	**.30**	**.30**	**.25**	**.46**
2. Cognitive Flexibility	**.28**	**.29**	**.25**	**.34**	**.50**
3. Achievement Orientation	**.27**	**.31**	**.43**	**.31**	**.37**
4. Fitness Motivation	**.16**	.10	**.23**	**.27**	**.14**
5. IS – Diplomacy	**.68**	**.44**	.09	**.20**	**.31**
6. Stress Tolerance	.10	-.09	**.22**	**.51**	.08
7. Hostility to Authority	.05	**-.30**	**-.33**	**-.33**	**-.14**
8. Self-Esteem	**.25**	**.20**	**.37**	**.37**	**.34**
9. Narcissism	**.13**	-.01	.02	**-.17**	.09
10. Cultural Tolerance	**.25**	**.41**	**.22**	**.25**	**.28**
11. Internal LOC	.09	**.18**	**.28**	**.37**	**.28**

Note. $n = 315$. Statistically significant correlations are bolded, $p < .05$ (two-tailed). Extr=Extraversion, Agr=Agreeableness, Cons=Conscientiousness, ES=Emotional Stability, Intel=Intellectance. Underlined correlations were hypothesized to be the highest in the row.

Faking Research

Next, we administered the RBI to another sample of new recruits to assess the fakability of the items and scales. This version of the RBI included 98 items retained from the pilot test RBI and 37 new items added to improve the psychometric properties of the existing scales. The .

intent was to eliminate items (and perhaps entire scales) that were easily fakable to create a self-report test suitable for use under operational conditions where the temptation to fake is high. Another critical consideration was to shorten the RBI so that test administration in the concurrent validation would take no longer than 30 minutes. However, one additional scale, measuring Gratitude Towards Others, was added out of concern that the RBI Cultural Tolerance scale might be an overly fakable measure of Agreeableness.

Sample and Design

The revised RBI was administered twice to a sample of 200 new recruits. All recruits initially completed the test under normal conditions in which they were asked to respond honestly (referred to as the Honest condition). Next, approximately half of these recruits were asked to retake the RBI while imagining that the results would affect their chances of joining the Army and getting the MOS they desired (referred to as the Fake Operational condition). The other half of recruits were asked to retake the RBI under the same instruction set as the Fake Operational Condition, but with additional explicit hints from the test administrators about how to maximize their RBI score (referred to as the Fake Operational with Coaching Condition).[28]

Analyses and Results

Statistical analyses were performed to identify RBI items and scales that were candidates for removal. First, items showing poor item/total scale correlations under the Honest condition were removed. Second, item correlations with the RBI Lie scale under the Honest condition were examined to reveal the degree to which each item may be contaminated with variance reflecting social desirability. Items correlating highly with the Lie scale were deleted. Third, correlations between the same item in the Honest and Fake Operational conditions were examined. The goal was to retain items with high Honest/Fake Operational correlations. The idea here was to retain items that resulted in similar rank orderings of respondents across conditions, with the assumption that using such items under operational conditions would result in a rank ordering of respondents that is similar to that achieved under an honest response condition.[29] Compared to faked scores, honest scores are more purely a reflection of construct-valid variance than variance arising from socially desirable responding. Thus, all else being equal, to the extent that "faked" item scores reflect variance due to socially responding, the correlation between "faked" and "honest" scores would be attenuated (assuming social desirability is relatively unrelated to the construct of interest). As such, we strived to retain items that correlated highly across conditions, with the idea that the variance they shared represented construct-valid variance.

---

[28] Exact instructions for the two faking conditions are provided in Appendix F. Because there are no plans to provide test takers with hints for scoring highly on the test should it become operational, it seems reasonable to conclude that the Fake Operational condition is more likely to create the variation in levels and types of faking that would occur under operational conditions (at least compared to responses provided under the Fake Operational with Coaching condition). Therefore, only the results obtained with the Fake Operational condition are presented herein.

[29] We acknowledge that the variation in item scores produced by differences in motivation observed in the operational setting will be different then the variation in item scores produced by the faking condition used here. Regardless of these differences, all else being equal, it is desirable to retain items that result in a rank ordering of respondents that is similar to that obtained under honest conditions despite the introduction of a contaminant source of variation into such items (i.e., variation arising from socially desirable responding).

In selecting items based on these criteria, some tradeoffs were required to balance these goals, yet still delete enough items to achieve the targeted RBI testing time. Upon removing items based on these criteria, 78 items remained on the RBI.

Table 8.4 presents descriptive statistics for the RBI scales in the faking research (after removal of the items described above), as well as the correlation of each scale with the RBI Lie scale in the Honest Condition and the same-scale correlation between the Honest and Fake Operational conditions. Scale internal consistency estimates are at or above the .60 level typically seen with rational biodata scales (Mumford & Owens, 1987). All but two scales have alphas higher than .70.

Table 8.4. Psychometric Statistics of RBI Scales in Faking Research

RBI Scale	# Items	Alpha	$r_{\text{Scale-Lie Scale}}$ in Honest Condition	$r_{\text{Scale-Scale}}$ Between Honest and Faking Conditions
1. Peer Leadership	6	.75	-.01	**.47**
2. Cognitive Flexibility	6	.67	.10	**.37**
3. Achievement Orientation	6	.68	**.24**	**.44**
4. Fitness Motivation	5	.75	.07	**.55**
5. Interpersonal Skills – Diplomacy	5	.79	**.15**	**.54**
6. Stress Tolerance	8	.70	**.16**	**.35**
7. Hostility to Authority	8	.72	-.11	**.50**
8. Self-Esteem	5	.78	.09	**.53**
9. Narcissism	6	.70	**.18**	**.57**
10. Cultural Tolerance	5	.71	**.14**	**.57**
11. Internal Locus of Control	8	.70	**.20**	**.50**
12. Gratitude	3	.57	**.19**	**.39**
13. Lie Scale	7	---	---	---

Note. n_{honest}= 186-200, n_{faking}= 94-100. Statistically significant correlations are bolded, $p < .05$ (two-tailed).

Examination of the fourth column in Table 8.4 reveals that only two RBI scales, Achievement Orientation and Internal Locus of Control, correlate .20 or higher with the Lie scale. Interpretation of the fifth column in Table 8.4 requires some caution. Previous experience with rational biodata scales in operational settings indicates that approximately 10% of test-takers trigger more than three faking items (Kilcullen et al., 2002; White, Gregory, Kilcullen, Galloway, & Nedegaard, 2001). In contrast, with the Fake Operational instructions over 47% of the subjects triggered more than three faking items, and 20% of the subjects triggered every faking item. This suggests that the Fake Operational instructions produce an artificially severe response set that overstates the degree to which RBI scores may be inflated in operational settings. With this caveat in mind, Table 8.4 reveals that seven of the 11 scales achieved Honest/Fake correlations of .50 or above, which is not much lower than the internal consistencies of the scales. Stress Tolerance and Cognitive Flexibility were most affected by explicit instructions to fake, with Honest/Fake correlations in the mid .30s.

Table 8.5 shows the effect sizes for Faking-Honest comparisons within each scale. By far the largest effect size was observed with the RBI Lie scale, indicating that this scale is sensitive to deliberate response distortion. Taken together with previous research demonstrating a high correlation between this scale and the Army Assessment of Background and Life Experiences (ABLE) Lie scale

(Kilcullen et al., 1995), the results support the construct validity of the RBI Lie scale. Aside from the RBI Lie scale, two biodata scales showed an effect size greater than 1.0, although all of the effect sizes were statistically significant. The median RBI scale effect size was 0.87.

When evaluating the "fakability" of the RBI scales, the relative order of respondents across Honest and Fake Operational conditions and effect size differences across conditions have different implications. For example, to the extent that the Fake Operational condition reproduces the variance in test taker responses that is attributable to response distortion in an operational context and the higher the Honest-Fake Operational correlations are, the more confident we can be that the scale will remain construct-valid when administered operationally. The results with respect to effect sizes speak not to the sources of variance in scale scores, but rather to the degree to which responses may be elevated in operational settings, and as such, the results have implications for setting cut-off scores in operational samples when a specific pass rate is desired.

Table 8.5. Honest-Operational Faking Effect Sizes of the RBI Scales

		Honest		Fake	
RBI Scale	d_{FH}	*M*	*SD*	*M*	*SD*
1. Peer Leadership	0.94	3.43	0.68	4.07	0.88
2. Cognitive Flexibility	0.87	3.54	0.61	4.07	0.74
3. Self-Esteem	0.98	3.79	0.63	4.41	0.66
4. Achievement Orientation	1.12	3.33	0.64	4.05	0.80
5. Fitness Motivation	1.04	3.11	0.79	3.93	1.02
6. Interpersonal Skills – Diplomacy	0.64	3.69	0.76	4.18	0.74
7. Stress Tolerance	0.91	2.85	0.58	3.38	0.72
8. Hostility to Authority	-0.87	2.42	0.63	1.87	0.67
9. Narcissism	0.40	3.42	0.65	3.68	0.78
10. Cultural Tolerance	0.70	3.81	0.67	4.28	0.72
11. Internal LOC	0.87	3.57	0.55	4.05	0.67
12. Gratitude	0.24	3.28	0.75	3.46	0.86
13. Lie Scale	2.40	0.07	0.15	0.43	0.40

Note. n_{honest}= 186-200, n_{faking}= 94-100. d_{FH} = Effect size for Faking-Honest comparisons. All effect sizes are statistically significant ($p < .05$).

Changes for the Field Test

An important goal for the field test version was to shorten the test so that very slow readers could still finish the instrument in 30 minutes. Although the pilot test and faking versions of the RBI both exceeded 130 items, a length of 100 items was targeted for the field test version to achieve this goal. When shortening the test, the primary consideration was to try to keep RBI scale alphas above .60 – a level generally considered acceptable for rational biodata scales (Mumford & Owens, 1987). In addition, an attempt was made to preserve the relatively low scale correlations with the Lie scale observed in Table 8.4 when shortening the test.

In the faking research, we identified 78 items to carry forward for use in the field test version of the RBI. In addition to these items, items for two additional constructs were added. During discussions among psychologists working on the Select21 project, it was noted that the Hostility to Authority scale might not inversely measure respect for authority, and that respect for authority might be a useful predictor of both Soldier performance and retention. Respect for

Authority was defined as valuing the opinions and advice of teachers and bosses. Four items were written to measure this construct and administered as part of the field test version of the RBI.

The other new RBI scale was developed to measure Army Identification, which is defined as the individual's intrinsic interest in becoming a Soldier. Once again, the hypothesis was that this construct would be a good predictor of both Soldier performance and attrition. Army Identification is similar to Meyer and Allen's (1991) concept of affective commitment, although affective commitment is measured for those with tenure in the organization. Among officers, affective commitment has been positively related to propensity to stay in the Army (Teplitsky, 1991).

Survey questions measuring Army affective commitment (Tremble, Payne, Finch, & Bullis, 2003) and military affective commitment (Heffner & Gade, 2003) were reviewed for possible modification as rational biodata items focusing on pre-enlistment personal identification with the Army. Three of these items were re-worked into rational biodata items relevant to new recruits. Four new biodata items were written to complete this scale. In a separate data collection, the Army Identification biodata scale was administered as part of a test battery to 155 enlisted Soldiers. Also administered was the Army affective commitment scale. The correlation between Army Identification and Army Affective Commitment was $r = .60$ ($p < .001$). With the addition of these items, the revised RBI that was administered during the field test consisted of 100 items.

Field Test

Sample

Data were gathered on 675 new recruits at Forts Knox and Jackson. A total of 71 cases were discarded. The most common reason for being discarded was a failure to complete enough of the test, although other cases were discarded because test administrators or item analyses revealed that respondents did not respond seriously to the test questions. The result was an analysis sample of 604 cases.

Analyses and Results

Descriptive statistics and intercorrelations of the RBI scales are presented in Table 8.6. A median scale alpha of .65 was obtained, and all but the shortened Stress Tolerance scale and the Gratitude scale achieved an internal consistency estimate of at least .60. An examination of the last row in Table 8.6 reveals that 11 out of 14 RBI scales were correlated less than .10 with the RBI Lie scale. This compares favorably with the observed correlations between the scales and the RBI Lie scale under Honest Conditions in the previous iteration of the RBI. Overall, the results generally suggest that the RBI was shortened in an adequate manner.

Examination of the correlation matrix in Table 8.6 reveals only one observed correlation at the .50 level or above (Self-Efficacy with Achievement Motivation). Generally speaking, these two scales showed the highest correlations with other scales in the shortened version of the RBI. Excluding Self-Efficacy and Achievement Motivation, scale intercorrelations were fairly low, with only three observed intercorrelations greater than .40. The moderate negative correlation between Hostility to Authority and Respect for Authority ($r = -.24$) supports the notion that these scales are not opposite ends of the same construct.

Table 8.6. Field Test of RBI Scales

Scale	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Peer Leadership	3.47	0.63	.66														
2. Cognitive Flex	3.49	0.61	**.47**	.71													
3. Achievement	3.51	0.55	**.49**	**.42**	.64												
4. Fitness Motivation	3.28	0.68	**.18**	**.14**	**.22**	.64											
5. Interp. Skills – Dipl.	3.58	0.74	**.44**	**.21**	**.34**	**.11**	.68										
6. Stress Tolerance	2.87	0.50	.02	**.14**	.03	**.12**	**.15**	.52									
7. Hostility to Authority	2.53	0.66	-.01	**-.13**	**-.24**	-.04	**-.10**	**-.34**	.65								
8. Self-Esteem	3.93	0.60	**.49**	**.39**	**.50**	**.28**	**.37**	**.11**	**-.18**	.68							
9. Cultural Tolerance	3.74	0.73	**.29**	**.38**	**.27**	**.12**	**.41**	**.27**	**-.36**	**.45**	.67						
10. Internal LOC	3.49	0.59	**.22**	**.25**	**.28**	**.13**	**.30**	**.31**	**-.37**	**.39**	**.30**	.68					
11. Army Identification	3.56	0.70	**.27**	**.14**	**.27**	**.26**	**.26**	**.18**	**-.16**	**.38**	**.27**	**.26**	.71				
12. Respect for Authority	3.45	0.69	**.18**	**.28**	**.43**	.07	**.12**	.04	**-.24**	**.24**	**.15**	**.16**	**.11**	.64			
13. Narcissism	3.71	0.55	**.37**	**.20**	**.36**	.05	**.21**	**-.23**	**.08**	**.43**	**.16**	**.14**	**.16**	**.10**	.61		
14. Gratitude	3.45	0.72	.03	**.09**	**.11**	.00	**.11**	.01	**-.18**	.05	**.17**	**.11**	.02	**.28**	-.03	.57	
15. Lie Scale	0.07	0.12	.00	.05	**.08**	.06	.03	**.16**	**-.13**	**.10**	.07	.04	.08	.03	-.04	**-.10**	---

Note. ***n*** = 604. Internal consistency estimates are in the diagonal. Statistically significant correlations are bolded, $p < .05$ (two-tailed).

The effect sizes of the RBI scales for gender and race are presented in Tables 8.7 and 8.8, respectively. Females tended to score lower than males in Fitness Motivation and Stress Tolerance but also lower in Hostility to Authority and higher in Achievement Orientation. Hispanics scored similarly to Whites on most RBI scales, although Hispanics were lower in Hostility to Authority and higher in Cultural Tolerance. The most marked difference was that Hispanics triggered more faking items than Whites. Whether this is the result of deliberate faking or some other effect (e.g., cultural differences, poor reading skills) might be a topic for future research.

Table 8.7. RBI Scores by Gender

		Male		Female	
RBI Scale	d_{FM}	*M*	*SD*	*M*	*SD*
Peer Leadership	-0.03	3.48	0.65	3.46	0.60
Cognitive Flexibility	0.06	3.48	0.64	3.52	0.55
Achievement Orientation	**0.37**	3.45	0.54	3.65	0.53
Fitness Motivation	**-0.40**	3.37	0.66	3.10	0.68
Interpersonal Skills – Diplomacy	**0.20**	3.54	0.76	3.69	0.71
Stress Tolerance	**-0.31**	2.92	0.48	2.77	0.54
Hostility to Authority	**-0.51**	2.62	0.63	2.30	0.66
Self-Esteem	0.13	3.90	0.63	3.98	0.52
Narcissism	**0.23**	3.67	0.57	3.80	0.51
Cultural Tolerance	**0.23**	3.69	0.74	3.86	0.69
Internal Locus of Control	**0.27**	3.44	0.60	3.60	0.54
Army Identification	0.00	3.56	0.69	3.56	0.74
Respect for Authority	**0.34**	3.38	0.70	3.62	0.65
Gratitude	0.05	3.44	0.74	3.48	0.65
Lie Scale	0.00	0.07	0.11	0.07	0.13

Note. n_{male}= 399-417, n_{female}= 177-183. d_{FM}= Effect size for Female-Male mean difference. Effect sizes calculated as (mean of non-referent group-mean of referent group)/*SD* of the total group. Referent groups (e.g., Males are listed second in the effect size subscript). Statistically significant effect sizes are bolded, p <.05 (two-tailed). A positive effect size indicates that in average the referent group performs better in the tests.

By far the largest Black/White difference was seen in the Army Identification scale, with Black recruits scoring on average one-half *SD* lower. Blacks also scored lower in Fitness Motivation and higher in Narcissism. On the other hand, they tended to score higher in Achievement Orientation, which has been a strong predictor of performance in other settings (Kilcullen et al., 1999, 2002). Perhaps Black recruits are more likely to view the Army as an opportunity for career advancement rather than being intrinsically interested in being a Soldier. Racial differences with respect to constructs like affective, normative, and continuance commitment have received relatively little attention in the literature. Karrasch (2003) examined White/Non-White differences, but separate analyses of White/Black and White/Hispanic comparisons were not performed. This might be an interesting topic for future research.

Table 8.8. RBI Scores by Race/Ethnic Group

			White		Black		White Non-Hispanic		Hispanic	
RBI Scale	d_{BW}	d_{HW}	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Peer Leadership	-0.02	0.07	3.48	0.59	3.47	0.73	3.47	0.58	3.51	0.75
Cognitive Flexibility	-0.11	0.11	3.48	0.62	3.55	0.55	3.47	0.63	3.54	0.54
Achievement Orientation	**0.27**	0.13	3.48	0.55	3.63	0.52	3.47	0.54	3.54	0.59
Fitness Motivation	**-0.27**	0.03	3.32	0.66	3.14	0.79	3.31	0.66	3.34	0.64
IS – Diplomacy	-0.10	0.15	3.60	0.73	3.53	0.82	3.59	0.73	3.70	0.68
Stress Tolerance	-0.20	-0.06	2.91	0.50	2.81	0.47	2.90	0.51	2.87	0.45
Hostility to Authority	0.06	**-0.27**	2.53	0.64	2.57	0.66	2.55	0.64	2.38	0.69
Self-Esteem	0.02	0.12	3.91	0.59	3.92	0.62	3.91	0.60	3.98	0.55
Narcissism	**0.33**	0.20	3.66	0.55	3.84	0.56	3.66	0.55	3.77	0.58
Cultural Tolerance	-0.13	**0.36**	3.73	0.72	3.64	0.83	3.71	0.73	3.97	0.64
Internal Locus of Control	-0.17	-0.12	3.52	0.59	3.42	0.54	3.53	0.59	3.45	0.56
Army Identification	**-0.50**	-0.11	3.61	0.70	3.26	0.68	3.62	0.72	3.54	0.56
Respect for Authority	0.19	-0.12	3.43	0.67	3.56	0.73	3.44	0.69	3.36	0.59
Gratitude	-0.07	-0.24	3.49	0.68	3.44	0.77	3.49	0.68	3.33	0.81
Lie Scale	0.08	**0.45**	0.07	0.12	0.08	0.13	0.06	0.11	0.11	0.14

Note. n_{white} = 343-407 n_{black} = 81-84, $n_{hispanic}$=71-73. d_{BW} = Effect size for Black-White mean difference. d_{HW} = Effect size for Hispanic-White Non-Hispanic mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., White) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

A principle components analysis using varimax rotation was performed on the scale scores to assess the dimensionality of the RBI. Examination of the scree plot yielded a 4-factor solution. The solution converged in six iterations and accounted for 58% of the variance. The RBI Peer Leadership, Achievement Orientation, and Self-Efficacy scales had the highest loadings on the first factor, called "Surgency." The second factor, labeled "Emotional Stability," received the highest loadings from the RBI Stress Tolerance, Hostility (negative loading), Cultural Tolerance, and Internal Locus of Control scales. The RBI Respect for Authority and Gratitude scales had high loadings on the third factor, labeled "Appreciation of Others," and the RBI Fitness Motivation and Army Identification scales loaded the strongest on the fourth factor, labeled "Rugged Orientation."

Preparing for the Concurrent Validation

Some RBI scales and items merited additional attention prior to the concurrent validation. The internal consistency of the Gratitude scale was unacceptably low, and contrary to expectations it was not a more fake resistant measure of Agreeableness compared to the Cultural Tolerance scale. Therefore the Gratitude scale was dropped from the RBI. The Stress Tolerance scale also had low internal consistency, but this construct was considered too important not to measure. Therefore, we replaced the Gratitude items with Stress Tolerance items to bolster the psychometric properties of the Stress Tolerance scale. We also replaced four other psychometrically weak items from various scales with previously discarded RBI items. These items, which had been administered during pilot testing, were subsequently shown to be related to early attrition.

CHAPTER 9: THE WORK SUITABILITY INVENTORY

Rodney A. McCloy and Dan J. Putka
HumRRO

Background

Researchers generally agree that people can fake self-report personality assessments (Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Ones, Viswesvaran, & Korbin, 1995) and that many will do so in operational selection settings (Hough, 1996, 1997, 1998; Rosse, Stechler, Miller, & Levin, 1998). Researchers disagree, however, regarding the extent to which faking affects the criterion-related validity of these assessments. Although many researchers have found that faking has little or no effect on criterion-related validity estimates (e.g., Barrick & Mount, 1996; Hough et al., 1990; Ones, Viswesvaran, & Reiss, 1996), other evidence suggests faking does change the rank-order of applicants in the upper tail of the distribution and results in the selection of individuals with lower-than-expected performance scores (Mueller-Hanson, Heggestad, & Thornton, 2003; Zickar, 2000). Given our experience with the Army's Assessment of Individual Motivation (AIM; Knapp, Waters, & Heggestad, 2002), we believe that response distortion poses a dauntingly high hurdle to the personnel selection specialist interested in using temperament measures in an operational setting.

Recent efforts to mitigate response distortion have centered on forced-choice formats. Although forced-choice formats have a demonstrated capacity to reduce the effects of faking (Jackson, Wrobleski, & Ashton, 2000; White & Young, 1998; Wright & Miederhoff, 1999), they result in ipsative response data (Hicks, 1970). Hicks defined ipsative scores as scores for which "each score for an individual is dependent on his own scores on other variables, but is independent of, and not comparable with, the scores of other individuals" (p. 167). Ipsative data allow a researcher to say such things as "David has higher standing on Conscientiousness than he does on Emotional Stability." The researcher cannot compare David's standing on Conscientiousness to the standing of any other person on that trait, however. For example, Maria might indicate a higher standing on Emotional Stability than on Conscientiousness, but this does not imply she has lower standing on the trait that David does.

One approach to reducing the ipsativity of a forced-choice measure involves introducing foil (i.e., dummy) constructs—constructs we do not wish to score. This approach reduces ipsativity in the responses because one can score relatively high or relatively low on all targeted constructs (when they are paired only with foil constructs).[30] Some ipsativity remains, however, because the forced-choice response depends upon the respondent's standing on the targeted and foil constructs in each pair.

Explaining this further, consider a measure comprising 10 statements, 1 for each of 10 constructs, with Conscientiousness the construct of interest and the 9 other constructs serving as foils. If one pairs the Conscientiousness statement with each of the nine dummy statements, then

[30] One cannot score high/low on all items in an ipsative measure. Rather, the scores of each item are at least partly conditional on the scores of other items. For example, if ranking stimuli, one must be ranked highest and one lowest. Indeed, with k items, once k-1 items have been ranked, the rank of the k^{th} item is fully determined.

a respondent can attain a Conscientiousness score ranging from 0 to 9 (i.e., the number of times the respondent selects the Conscientiousness statement as "More like me" when it appeared with a statement assessing a foil construct). Neither the total score nor the variance for the Conscientiousness score suffers from intra-measure constraints (i.e., the Conscientiousness score need not suffer from its statement appearing with statements reflecting other target constructs). Nevertheless, the total score does not provide purely normative trait information, because the number of times the respondent selects the Conscientiousness statement as "More like me" necessarily depends on the set of foil constructs assessed by the statements paired with the Conscientiousness statement in each paired comparison. That is, a respondent might obtain a Conscientiousness score of 7 when responding to a measure containing constructs A through I but score only 3 when responding to a measure containing constructs J through R. Thus, although ipsativity fades, it does not exit the stage entirely. One will likely attain better approximations of normative construct standings, however, to the extent that one more fully samples the content space of interest (here, personality traits).

Development Process and Scoring Plan

Development

The Work Suitability Inventory (WSI)[31] comprises 16 statements (stems) that describe work requirements. All but one of the statements come from the Work Styles portion of the O*NET content model (Borman, Kubisiak, & Schneider, 1999), although we have simplified their wording to make them more accessible to recruits (see Figure 9.1).[32,33] Basing the measure on the O*NET Work Styles taxonomy aligns this portion of the measurement development activity with the person-environment (P-E) fit research in the project (see Chapter 13 of this report). It also serves as an additional deterrent to prevent respondents from gaming their answers, because all stems were written to be of comparable social desirability. Finally, the O*NET Work Styles taxonomy provides the WSI with a defensible taxonomic base upon which to argue that the stems from target traits appear with an appropriate set of dummy traits. This is important, because we believe that the WSI will be most useful and informative (and have the best chance for demonstrating predictive validity) when respondents base their rankings on as full a range of traits as possible (rather than just those traits appearing in the Select21 KSAs). For these reasons, the WSI statements retain the O*NET Work Style labels rather than adopting the Select21 KSA labels.[34]

---

[31] The WSI was originally entitled the Person-Organization Personality (POP) Hybrid (cf. McCloy, Putka, Van Iddekinge, & Kilcullen, 2003). We now apply the term to a type of measure rather than to a specific measure. For example, the WSI could be described as a hybrid of a person-organization (PO) fit measure and traditional personality (P) assessment (thus, the term "POP" hybrid).
[32] Some items underwent multiple re-wordings.
[33] The statement not taken from the O*NET addresses cultural tolerance.
[34] This also obviates the need to devise a label for Leadership Orientation, Persistence, and Initiative—traits that appear in the Work Styles taxonomy but not in the Select21 KSA listing (see Chapter 1).

Current WSI Wording	Original O*NET Work Styles Wording
Work that requires ... showing a cooperative and friendly attitude towards others I dislike or disagree with. (Agreeableness)	Job requires being pleasant with others on the job and displaying a good-natured, cooperative attitude encourages people to work together (Cooperation)
Work that requires ... being open to change (positive or negative) and a lot of variety. (Intellectance)	Job requires being open to change (positive or negative) and to considerable variety in the workplace (Adaptability/Flexibility)
Work that requires ... leading, taking charge, and giving direction. (NA)	Job requires a willingness to lead, take charge, and offer opinions and direction (Leadership Orientation)
Work that requires ... accomplishing tasks alone, with little supervision or help from others. (Self-Reliance)	Job requires developing own ways of doing things, guiding oneself with little or no supervision, and depending mainly on oneself to get things done (Independence)
Work that requires ... setting challenging goals and working continuously to attain them. (Achievement Motivation)	Job requires establishing and maintaining personally challenging achievement goals and exerting effort toward task mastery (Achievement/Effort)
Work that requires ... consistently meeting obligations and completing duties on time. (Dependability)	Job requires being reliable, responsible, and dependable, and fulfilling obligations (Dependability)
Work that requires ... dealing effectively with high-stress situations and accepting frequent criticism. (Emotional Stability)	Job requires accepting criticism, and dealing calmly and effectively with high-stress situations (Stress Tolerance)
Work that requires ... much creativity and original thinking to perform successfully. (Intellectance)	Job requires creativity and alternative thinking to come up with new ideas for and answers to work-related problems (Innovation)
Work that requires ... maintaining composure and keeping emotions and behavior in check even in very difficult circumstances. (Emotional Stability)	Job requires maintaining composure, keeping emotions in check even in very difficult situations, controlling anger, and avoiding aggressive behavior (Self-Control)
Work that requires ... being sensitive to others' needs and feelings and being understanding. (Social Perceptiveness)	Job requires being sensitive to others' needs and feelings and being understanding and helpful on the job (Concern for Others)
Work that requires ... high levels of energy and stamina to perform successfully. (Potency)	Job requires the energy and stamina to accomplish work tasks (Energy)
Work that requires ... working closely with others (instead of alone) to get tasks completed. (Team Orientation, Affiliation)	Job requires preferring to work with others rather than alone and being personally connected with others on the job (Social Orientation)
Work that requires ... being thorough and paying close attention to details. (Dependability)	Job requires being careful about detail and thorough in completing work tasks (Attention to Detail)
Work that requires ... performing tasks that take a long time to "get right" and overcoming several obstacles along the way. (NA)	Job requires persistence in the face of obstacles on the job (Persistence)
Work that requires ... taking on new or additional responsibilities that may fall outside of my job duties. (NA)	Job requires being willing to take on job responsibilities and challenges (Initiative)
Work that requires ... interacting with people of different cultures and backgrounds, and appreciating differences in their values, opinions, and beliefs (Cultural Tolerance)	NA[a]

Note. The target Select21 KSA appears in parentheses following each WSI statement; the O*NET work style appears in parentheses following each O*NET statement. NA = Not applicable. [a]No O*NET work style assesses Cultural Tolerance.

*Figure 9.1. The Work Suitability Inventory statements and their original O*NET versions.*

The WSI attempts to distract respondents from thinking about how best to game their answers to a temperament assessment by redirecting their thoughts toward P-E fit. The current version of the WSI presents respondents with a computerized card-sorting task. Specifically, the computer displays 16 cards on the screen. Each card contains one of the work characteristic statements and an identifying letter ranging from "A" to "P." Respondents must "sort the 16 cards in terms of how well you think you would perform the type of work described by the cards. Cards containing types of work that you think you would perform best should be ranked highest; cards containing types of work that you think you would perform worst should be ranked lowest." Respondents sort the 16 cards by using the computer mouse to drag and drop the cards into 16 boxes outlined on the screen. The card sorted into the first box describes the type of work the respondent believes s/he would perform best; the card sorted into the last box describes the type of work the respondent believes s/he would perform least well.

Scoring Plan

There are multiple options for scoring the WSI, depending on whether we want to use it for traditional personality assessment applications or P-E fit applications. In the sections below, we briefly describe two of these options.

Option 1: Scoring Target Constructs Only

Under this scoring option, which treats the WSI as a means for conducting traditional personality assessment, we would score only those WSI statements assessing target constructs; the remaining statements would serve as foil (dummy) constructs (i.e., constructs we decide not to target). Comparing the target constructs to foil constructs reduces score ipsativity, and the reduction varies proportionally with the number of foil constructs. Constructs selected as target constructs will be those hypothesized to be most related to the criterion of interest (e.g., attrition, job performance, re-enlistment). Target constructs might vary for different criteria; indeed, this is an explicit goal of the WSI, because the criterion-based scoring makes it harder to fake the instrument maximally for all its possible uses (e.g., predicting job performance, predicting attrition).

To illustrate how this scoring strategy might work, consider the following example. Assume we select five WSI constructs as predictors of a given criterion variable (i.e., as target constructs). The score for each target construct would be its rank relative to the foil constructs. That is, targets would receive a score of "(F+1) – $rank_f$", where F is the number of foil constructs and $rank_f$ is the rank the target construct receives *among the F+1 statements* (the F foil constructs and the target construct in question). This means that the score for a target construct can range from a high of F to a low of 0. Returning to our example, we designate 5 of the 16 constructs as targets, leaving 11 traits as foils. Therefore, F+1 = 12 and the score for Target 1 equals 12 - $rank_f$. If Target 1 were ranked higher than any other construct (target or foil), $rank_f$ would equal "1" and the score for Target 1 would be 12 –1 = 11. If Target 1 were ranked third among all constructs, and the two constructs ranked higher were foils, $rank_f$ would equal 3 and Target 1 would receive a score of 12 – 3 = 9. If, however, the two constructs ranked higher than Target 1 were *other targets*, then $rank_f$ would again equal 1 (because $rank_f$ gives the rank of the target construct relative only to the foils—not to the other target constructs), and Target 1 would receive a score of 12 – 1 = 11. Note that in this latter situation, a score of 11 would also be

assigned to the two target constructs that were ranked first and second *overall* (i.e., ahead of Target 1), because they also receive higher ranks than any of the foils (i.e., they also have rank_f values of "1"). Were the data totally ipsative, different constructs could not receive the same score; thus, the data are only partially ipsative, thereby improving their statistical characteristics.

A simple variation on the above approach that we could also evaluate during subsequent efforts (e.g., concurrent validation, attrition database analyses) would be to use a set of foils for each target construct, that either (a) maximize the target construct's relation to the criterion of interest (i.e., maximize criterion-related validity), (b) maximize its correlation with an alternative measure of the given construct (i.e., maximize construct validity), or (c) some combination of these strategies. Thus, instead of using all available foils, we would select foils based on psychometrically valued criteria.

Option 2: Scoring All Constructs

Under this option, where the WSI can be used as a tool for assessing P-E fit, scores of "17 – rank" are assigned to *all* WSI statements. This results in complete ipsativity. Nevertheless, it gives us a rank ordering of each individual's perceived strengths when it comes to temperament-related requirements of work, as well as comprehensive coverage of the personality domain. As we discuss in Chapter 13, recruits' WSI profiles based on this scoring option will be correlated with an "environment-side" profile of the temperament-related requirements of Army work to assess recruits' fit to the Army environment.[35]

Managing Response Distortion

Given the problem of response distortion on self-report temperament measures, it is important to highlight how the WSI (through its scoring, design, and delivery) attempts to circumvent this problem. Respondents completing the WSI in an operational setting could try to distort their rankings to what they think the ideal personality for their desired MOS or the Army in general. The inclusion of "dummy" personality constructs (i.e., foils) that possess equal levels of social desirability (in terms of a selection application) is designed to mitigate this problem.

Furthermore, although respondents might try to distort the rank ordering of stems to match the ideal personality for a given MOS or the Army, such distortion might not detract from the criterion-related validity of the resulting score. Indeed, this particular form of distortion would indicate familiarity with the requirements of the Army work environment and realistic expectations about what the work requires. The literature on realistic job previews suggests that familiarity with the work environment and realistic expectations would contribute to criterion-related validity when predicting alternative criteria such as job satisfaction and attrition (Wanous, 1992). Thus, although this type of response distortion represents a source of contamination in WSI scores, it could very well serve as criterion-related contamination and thus enhance criterion-related validity.

[35] As part of the criterion field test, we collected data from NCOs regarding the temperament-related requirements of first-term Soldiers' work. These data, as well as their similarity to recruits' WSI profiles, are discussed in Chapter 13.

In addition, we can select which WSI constructs to treat as targets and which to treat as foils depending on the criteria of interest. Thus, for criterion Y_1, Achievement/Effort, Energy, and Leadership Orientation might serve as the target constructs, with the other 13 constructs serving as foils. Criterion Y_2, on the other hand, might require Innovation, Analytic Thinking, Stress Tolerance, and Energy as the keyed traits. This flexibility in how we treat constructs tapped by the WSI has great value for two additional reasons. First, the Army often desires to use the same instrument to predict a variety of criteria (e.g., using AIM to predict NCO performance, recruiter performance, and first-term attrition). Second, to the extent that we can convince respondents completing the WSI that the Army will use the results for a variety of purposes (thus another reason for covering the domain of personality), it may prevent the respondents from attempting to fake toward a given profile or in a certain direction. This latter point speaks to the importance of carefully crafting the instructions given to recruits so as to manage their frame of reference when completing the WSI.

Data Collections and Results

We have three main sources of WSI data at present: (a) pilot tests conducted in the fall of 2003, (b) faking research conducted in January and February of 2004, and (c) field tests conducted in the fall of 2004. In each of these data collections, the WSI was administered to new Army recruits as they processed through their reception battalions. Future WSI data will be obtained during the concurrent validation. This section presents the key results from each of the three data collections to date.

Pilot Test

Prior to administration of the WSI to new recruits as part of the pilot test, an initial paper-and-pencil version comprising 105 paired comparisons was "pre-piloted" on a sample of 177 Soldiers at AIT schoolhouses.[36] Soldiers were asked to select the one statement out of each pair that described the type of work they believed they "would be more successful at." Not surprisingly, Soldiers reacted quite negatively to the redundancy of the measure and sheer drudgery of the exercise. In addition, the measure required an inordinate amount of administration time (approximately 45 minutes). We therefore put this version aside in favor of an alternative response format that was administered to new recruits as part of the pilot test at the reception battalions.

The alternative version involved a manual card-sorting exercise and is essentially a paper-and-pencil version of the current computerized WSI. Recruits received a bundle of 16 cards and a sorting sheet containing 16 boxes. Each card contained one of the work characteristic statements and an identifying letter ranging from "A" to "P." Recruits were to "sort the 16 cards in terms of how well you think you would perform the type of work described by the cards. Cards containing types of work that you think you would perform best should be ranked highest; cards containing types of work that you think you would perform worst should be ranked lowest." Recruits indicated their choices on a scantron sheet.

[36] This original version did not yet include the Cultural Tolerance stem. It therefore comprised 15 statements that yielded (15*14)/2 = 105 paired comparisons.

For the pilot test, we administered the card-sort version of the WSI (WSI-CS) to new recruits at Forts Benning, Jackson, and Knox. A total of 331 recruits completed the measure. Data screening yielded a total of 310 valid cases.

The primary finding from the pilot test analyses involved correlations among the 16 WSI construct scores and scales from other measures that were administered as "markers." The marker measures, which were administered to recruits along with the WSI, helped us assess the degree to which the WSI statements reflected the constructs they were designed to tap.[37] The 50-item Likert scale measure of the Big Five personality traits from the International Personality Item Pool (IPIP) served as one marker measure. The IPIP is described further in Chapter 8.

Given the ipsative nature of the WSI construct scores, we also administered a Likert version of the WSI (the WSI-Likert). The WSI-Likert presented recruits with each of the 16 statements used on the WSI-CS and asked them to indicate the extent to which they agreed they could perform each of them well. Recruits made these ratings on a scale ranging from *Strongly Disagree* (1) to *Strongly Agree* (5). Comparing the magnitude of correlations between (a) each WSI measure and (b) the IPIP marker provides an indication of the extent to which ipsativity in the WSI-CS might attenuate the correlations observed with the IPIP marker measure. Table 9.1 shows descriptive statistics and intercorrelations for the WSI-Likert scales administered during the pilot test.[38]

To assess convergent validity of the WSI construct scores, we examined their correlations with the IPIP scale scores. Significant correlations between WSI construct scores and the IPIP scores to which they should theoretically be related provides evidence of convergent validity. Table 9.2 shows correlations of the IPIP with the WSI-CS and WSI-Likert construct scores. The pattern of correlations indicates reasonable convergent validity for the WSI constructs. For example, the WSI constructs Achievement/Effort, Attention to Detail, and Dependability—all subfactors of Conscientiousness—correlate highly with the IPIP scale Conscientiousness, although Achievement/Effort shows slightly higher correlations with IPIP scales Emotional Stability and Intellectance. Similarly, WSI constructs Self-Control and Stress Tolerance correlate most highly with IPIP Emotional Stability (while exhibiting high correlations with Conscientiousness, as well). Other constructs (e.g., Concern for Others and Cooperation) also show desirable correlational patterns.

One thing to note is that the WSI-Likert correlates more highly with the IPIP than does the WSI-CS. Although the latter correlations might seem low, the ipsativity of the rank data attenuates their magnitude. The results, therefore, are encouraging.

[37] For this analysis, no constructs served as foils. Rather, trait scores for the WSI were calculated using the second scoring option described earlier ("17-rank"). Hence, the WSI data were fully ipsative.

[38] Note that no internal consistency reliability estimates for WSI-Likert scores are shown in Table 9.2. This is because each trait on the WSI-Likert was assessed with a single-item. Although single-item measures of traits are notoriously unreliable, our intent here was to replicate the WSI-CS in a Likert format. The implications of using these single-item WSI-Likert scores for interpreting observed correlations will be discussed in subsequent sections.

Table 9.1. Descriptive Statistics and Intercorrelations for WSI-Likert Scores

WSI Trait	*M*	*SD*	*Skew*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1 Achievement/Effort	3.86	0.93	-0.67	-														
2 Adaptability/Flexibility	3.77	0.89	-0.85	**.23**	-													
3 Attention to Detail	3.89	0.94	-0.82	**.43**	**.25**	-												
4 Concern for Others	3.71	1.03	-0.64	**.15**	**.28**	**.12**	-											
5 Cooperation	3.39	1.12	-0.51	**.20**	**.37**	**.20**	**.32**	-										
6 Dependability	3.87	0.94	-0.77	**.54**	**.29**	**.51**	**.19**	**.27**	-									
7 Energy	3.90	0.98	-0.77	**.39**	**.30**	**.30**	**.17**	**.17**	**.33**	-								
8 Independence	3.75	1.06	-0.70	**.19**	.05	**.10**	-.07	-.07	**.12**	**.08**	-							
9 Initiative	3.60	0.97	-0.46	**.29**	**.24**	**.34**	.06	**.22**	**.30**	**.16**	**.09**	-						
10 Innovation	3.79	1.01	-0.58	**.19**	**.22**	**.29**	**.21**	**.08**	**.19**	**.18**	**.14**	**.15**	-					
11 Leadership Orientation	3.72	1.04	-0.64	**.23**	**.20**	**.34**	**.08**	**.14**	**.29**	**.27**	**.17**	**.19**	**.28**	-				
12 Persistence	3.57	1.06	-0.47	**.44**	**.28**	**.41**	**.14**	**.20**	**.38**	**.25**	**.17**	**.44**	**.27**	**.18**	-			
13 Self-Control	3.88	0.93	-0.63	**.34**	**.20**	**.37**	**.12**	**.25**	**.33**	**.26**	**.11**	**.29**	**.22**	**.27**	**.33**	-		
14 Social Orientation	3.77	1.04	-0.70	**.22**	**.29**	**.22**	**.36**	**.34**	**.26**	**.31**	**-.23**	**.16**	**.18**	**.25**	**.17**	**.16**	-	
15 Stress Tolerance	3.40	1.10	-0.44	**.23**	**.23**	**.29**	.04	**.24**	**.24**	**.20**	.05	**.30**	**.14**	**.20**	**.32**	**.40**	**.15**	-
16 Cultural Tolerance	3.95	1.02	-0.83	**.24**	**.34**	**.29**	**.35**	**.37**	**.24**	**.24**	-.04	**.23**	**.27**	**.20**	**.23**	**.25**	**.39**	**.17**

Note. $n = 506\text{-}520$. Boldface correlations are statistically significant ($p < .05$, one-tailed).

Table 9.2. Convergent Validity Estimates of the WSI with the IPIP Big Five Marker Measure

	IPIP Scale									
	Extraversion		Agreeableness		Conscient.		Emotional Stability		Intellectance	
WSI Construct	CS	Likert	CS	Likert	CS	Likert	CS	Likert	CS	Likert
Achievement/Effort	-.03	.03	.01	.06	.09	**.36**	**.11**	**.19**	**-.10**	.09
Adaptability/Flexibility	-.01	**.23**	.01	**.32**	**-.17**	**.24**	-.07	**.23**	**-.10**	**.16**
Attention to Detail	**-.13**	.07	**-.15**	**.18**	**.22**	**.53**	.05	**.23**	.01	**.23**
Concern for Others	.07	**.18**	**.29**	**.61**	-.08	**.11**	**-.19**	.05	-.08	**.17**
Cooperation	.00	.08	**.18**	**.29**	**-.14**	**.21**	-.01	**.27**	**-.11**	.09
Dependability	-.09	.07	-.05	**.24**	**.20**	**.58**	.06	**.21**	**-.15**	.09
Energy	**.10**	**.24**	-.03	**.13**	.08	**.23**	-.01	**.10**	.01	**.16**
Independence	**-.19**	-.07	**-.25**	-.09	-.03	**.14**	**-.13**	.08	-.03	**.11**
Initiative	**-.10**	.03	**-.10**	.07	-.02	**.26**	.03	**.21**	.00	**.11**
Innovation	.07	**.21**	.02	**.21**	**-.10**	**.13**	.02	**.11**	**.31**	**.61**
Leadership Orientation	**.30**	**.36**	.08	**.23**	**.10**	**.34**	.03	**.16**	**.23**	**.37**
Persistence	**-.13**	-.02	**-.13**	.02	.09	**.28**	.05	**.13**	.04	**.13**
Self-Control	**-.11**	-.03	**-.17**	.04	.00	**.31**	.09	**.31**	.05	**.16**
Social Orientation	.06	**.33**	**.12**	**.38**	**-.16**	**.11**	-.03	**.20**	**-.11**	**.17**
Stress Tolerance	.03	.08	-.09	.07	.05	**.20**	**.11**	**.22**	-.01	.07
Cultural Tolerance	**.10**	**.21**	**.21**	**.38**	-.08	**.20**	-.07	**.18**	.02	**.24**

Note. $n = 295$. Convergent validity estimates are zero-order Pearson correlations between the WSI constructs and IPIP scales. CS = Card-sort version of WSI. Likert = Likert version of WSI. Bolded correlations are statistically significant ($p < .05$, one-tailed).

To further address the issue of ipsativity in the WSI-CS, we calculated "same-construct" correlations between the card-sort and Likert-versions of the WSI (see Table 9.3). These correlations provide an indication of the degree to which trait scores on the WSI-CS provide approximations of recruits' normative standing on the traits assessed by the WSI. All correlations between corresponding WSI-CS and WSI-Likert were statistically significant in the positive direction. The average same-construct correlation was .37, with correlations ranging from a low of .20 (Adaptability/Flexibility) to a high of .54 (Leadership Orientation). Given the WSI-Likert trait scores are based on single-item measures of these traits, it is likely that these correlations are attenuated (i.e., they make it appear ipsativity has more of an effect on WSI-CS scores than it does). For example, on the IPIP, single-item IPIP reliability estimates ranged from .24 (Intellectance) to .41 (Extraversion) ($M = .33$). If the reliability of the WSI-Likert trait scores were in this range, it would suggest that corrected correlations between corresponding WSI-CS and WSI-Likert trait scores would be far higher than the observed correlations presented in Table 9.3. Taken together, the results in Tables 9.2 and 9.3 suggest that the WSI-CS trait scores are providing reasonable approximations of recruits' normative standing on the traits they were designed to measure.[39]

[39] Note that an additional factor that may be influencing the correlation between the card-sort and Likert versions of the WSI traits is correlated error. Specifically, because recruits were administered the WSI-CS and WSI-Likert during the same data collection session, it is possible that transient error associated with each measure (i.e., occasion-specific error) is inflating the level of correlation between the WSI-CS and WSI-Likert scales.

Table 9.3. Same-Construct Correlations between Card-Sort and Likert Versions of the WSI

WSI Construct	*r*	WSI Construct	*r*
Achievement/Effort	.38	Initiative	.35
Adaptability/Flexibility	.20	Innovation	.46
Attention to Detail	.29	Leadership Orientation	.54
Concern for Others	.44	Persistence	.33
Cooperation	.32	Self-Control	.31
Dependability	.24	Social Orientation	.34
Energy	.44	Stress Tolerance	.45
Independence	.48	Cultural Tolerance	.41

Note. n = 301. All correlations are statistically significant ($p < .05$, one-tailed).

Faking Research

To determine the degree to which (and the manner in which) respondents *could* distort their responses on the non-cognitive predictors developed for the Select21 project, faking research was conducted during the early months of 2004.[40] Recruits completed the WSI under an honest condition (n = 194) and one of two faking conditions. The first faking condition was a "fake maximum" condition where recruits were asked to respond in a way that would make them look as good to the Army as possible without fear of detection (n = 98). The second faking condition was a "fake maximum/avoid detection" condition where recruits were asked to look as good to the Army as they possibly could, but to do so in a way that would not make it look obvious they were trying to distort their responses (n = 99). (Complete instructions given to recruits in these faking conditions are provided in Appendix F.)

Table 9.4 shows means and standard deviations for WSI construct scores in each administration condition, as well as standardized effect sizes indexing the difference in scores across conditions.[41] The largest inflation of construct scores from honest to faking conditions occurred for Stress Tolerance ($d_{FM\text{-}H}$ = 0.99, $d_{FM/AD\text{-}H}$ = 0.89) and Dependability ($d_{FM\text{-}H}$ = 0.72, $d_{FM/AD\text{-}H}$ = 0.57). These results suggest that, on average, recruits thought that inflating their scores on Stress Tolerance and Dependability relative to the other constructs would make them look more attractive to the Army. Conversely the largest deflation of construct scores from honest to faking conditions occurred for Innovation ($d_{FM\text{-}H}$ = -0.81, $d_{FM/AD\text{-}H}$ = -0.71), Concern for Others ($d_{FM\text{-}H}$ = -0.77, $d_{FM/AD\text{-}H}$ = -0.67), and Independence ($d_{FM\text{-}H}$ = -0.62, $d_{FM/AD\text{-}H}$ = -0.71). These results suggest that on average, recruits thought that deflating their scores on these constructs would make them look more attractive to the Army. Interestingly, although elevation differences were apparent between honest and faking conditions, with the exception of Concern for Others, there were only minimal differences in standard deviations across conditions. Thus, it appears the WSI construct scores were still able to differentiate among recruits in the faking conditions.

[40] Given we examined faking under experimentally controlled conditions, we were primarily interested in assessing the degree to which recruits had the ability to fake the WSI and how they did so, rather than assessing the degree to which the WSI would be faked in an operational setting (a function of both the ability *and* the motivation to fake).
[41] Given that no criteria are yet available for informing the scoring of the WSI, trait scores reported in this section were calculated by taking the ranking each recruit gave to a WSI statement, and subtracting it from 17 (scoring option 2 discussed earlier). Thus, WSI trait scores range from 1 to 16, with higher scores being indicative of types of work that recruits thought they would perform best.

These findings led us to pose another question. Specifically, if variation was maintained across conditions, was the nature of that variation the same? If so, then recruits would be rank-ordered similarly across conditions. In a directed faking study, motivations to fake are likely equalized far more than they are in practice, and they therefore should have little impact on the rank ordering of respondents across conditions. Thus, a lack of relation between honest and faked scores should reflect (a) differences in ability to fake effectively, and/or (b) differences in compliance with the instruction sets. To investigate these possibilities, we examined the correlation between respondents' honest and faked scores (see the r columns in Table 9.4). The correlations between honest and faked WSI construct scores were generally low (e.g., r = -.36 to .25 for honest-fake max; r = -.14 to .23 for honest-fake max/avoid detection). These results suggest that there are notable individual differences in recruits' ability to fake the WSI effectively and/or their compliance with the faking instructions.

Finally, note the last row in Table 9.4. As mentioned earlier, one option for using recruits' WSI data would be to calculate correlations between each recruit's WSI profile and a profile that reflects the temperament-related requirements of Army work. The last row in Table 9.4 provides descriptive statistics for such a statistic calculated across conditions. Specifically, we calculated a Spearman rank-order correlation that reflects the similarity of recruits' WSI profiles to a profile based on NCOs' completion of the *Work Styles Supply Survey* (WSSS; see Chapter 13). In short, the WSSS was designed to generate a single profile (based on mean NCO rankings) that reflected how well each WSI construct described work performed by first-term Soldiers. Thus, to the extent that correlations between recruits' WSI profiles and the WSSS profile are higher in faking conditions than in the honest condition, recruits can be deemed capable of faking a profile that resembles the Army work required. Results presented in the last row of Table 9.4 show that the similarity of recruits' WSI profiles to the WSSS profile increased substantially from honest to faking conditions ($d_{FM\text{-}H}$ = 1.19, $d_{FM/AD\text{-}H}$ = 1.08). Nevertheless, it is important to note that the mean correlation between recruits' profiles and WSSS profile remained low in the faking condition (M_{FM} = .28, $M_{FM/AD}$ = .25). Furthermore, despite the increased profile similarity, the standard deviations in the faking conditions were actually slightly higher than in the honest condition. Hence, although recruits (on average) were able to fake a profile that was more similar to the Army profile when asked to do so, recruits varied markedly in their ability to do so.

Field Test

Sample

A total of 665 recruits completed the computerized version of the WSI as part of the predictor field test. Of the recruits with WSI data, 35 were removed from the sample because of various problems with their data (e.g., completed the measure in an unreasonably short period of time [less than 2 minutes and 20 seconds], ranked the cards alphabetically from A-P).

Descriptive Statistics

Table 9.5 shows descriptive statistics for each WSI construct score. As with the faking research data, construct scores were created by taking the ranking each recruit gave to a WSI statement, and subtracting it from 17. Thus, WSI construct scores range from 1 to 16, with higher scores being indicative of types of work that recruits' thought they would perform best.

Table 9.4. Descriptive Statistics and Fake-Honest Differences on WSI Constructs

Construct	Honest (H)		Fake Max (FM)		Fake Max / Avoid Detection (FM/AD)		FM-H		FM/AD-H		FM-FM/AD
	M	*SD*	*M*	*SD*	*M*	*SD*	*d*	*r*	*d*	*r*	*d*
Achievement/Effort	10.51	4.35	10.01	4.19	10.28	4.67	-0.12	-.04	-0.05	.19	-0.06
Adaptability/Flexibility	8.89	4.35	7.36	4.10	8.05	4.27	-0.35	-.25	-0.19	.02	-0.16
Attention to Detail	9.07	4.20	11.05	4.38	10.91	3.90	0.47	.12	0.44	.23	0.04
Concern for Others	8.13	5.03	4.25	3.85	4.78	3.99	-0.77	.05	-0.67	.14	-0.13
Cooperation	6.97	4.34	7.54	4.19	7.52	4.00	0.13	-.36	0.13	.07	0.01
Dependability	8.56	4.06	11.47	4.06	10.88	3.82	0.72	.00	0.57	.15	0.15
Energy	8.98	4.40	10.38	4.14	11.01	4.09	0.32	-.03	0.46	-.14	-0.15
Independence	8.72	4.92	5.68	4.40	5.21	4.12	-0.62	-.08	-0.71	.05	0.11
Initiative	7.59	3.59	8.38	4.03	8.78	4.13	0.22	-.10	0.33	.08	-0.10
Innovation	9.59	4.79	5.73	3.90	6.16	4.16	-0.81	-.06	-0.71	-.06	-0.11
Leadership Orientation	10.23	4.74	11.09	4.71	10.90	4.38	0.18	-.06	0.14	.10	0.04
Persistence	7.21	4.15	7.79	3.77	7.40	4.08	0.14	-.19	0.05	-.11	0.10
Self-Control	7.93	4.33	8.43	3.66	8.41	3.87	0.11	-.07	0.11	-.14	0.00
Social Orientation	8.98	4.52	8.89	4.30	8.41	4.58	-0.02	-.03	-0.12	.06	0.11
Stress Tolerance	5.53	4.44	9.95	4.19	9.49	4.52	0.99	.25	0.89	-.01	0.10
Cultural Tolerance	9.04	4.89	7.99	4.48	7.80	4.49	-0.22	-.14	-0.26	.18	0.04
WSI-WSSS Spearman *r*	-0.02	0.25	0.28	0.30	0.25	0.29	1.19	.17	1.08	.06	0.09

Note. $n = 179$ (honest); 91 (fake max); and 99 (fake max/avoid detection). d = standardized mean difference, which was calculated by subtracting the honest mean from the faking mean and dividing that difference by the standard deviation from the honest condition. For FM - FM/AD d values, the pooled standard deviations of fake max and fake max/avoid detection scores were used as divisors. r = zero-order correlation between honest and faking dimension scores. FM-H and FM/AD-H d and r statistics are based on within-subjects data. WSI-WSSS Spearman r = Spearman rank-order correlation indexing the similarity of construct ranks between each recruit's profile and the mean *Work Styles Supply Survey* (WSSS) profile (discussed in text).

Table 9.5. Descriptive Statistics for the WSI in the Field Test Sample

WSI Construct	*M*	*SD*	*Skew*
Achievement/Effort	10.45	4.60	-0.48
Adaptability/Flexibility	9.08	4.32	-0.19
Attention to Detail	9.37	4.39	-0.23
Concern for Others	7.78	4.83	0.19
Cooperation	8.43	4.42	-0.08
Dependability	9.11	4.02	-0.21
Energy	9.44	4.50	-0.25
Independence	7.77	5.03	0.14
Initiative	7.51	3.78	0.26
Innovation	9.07	4.69	-0.01
Leadership Orientation	9.40	4.52	-0.16
Persistence	7.21	4.19	0.52
Self-Control	8.16	4.29	0.14
Social Orientation	8.82	4.67	-0.04
Stress Tolerance	5.72	4.46	0.67
Cultural Tolerance	8.65	4.85	-0.13

Note. $n = 630$.

Recruits indicated they would be most effective at work that required Achievement/ Effort, Energy, Leadership Orientation, and Attention to Detail. Recruits indicated they would be least effective at types of work that required Stress Tolerance, Persistence, and Initiative. Examination of skew statistics and score distributions (see Table 9.6) revealed that most of the WSI construct scores were normally distributed. Exceptions to this were Stress Tolerance and Persistence (which were moderately positively skewed), and Achievement/Effort (which was moderately negatively skewed).

Table 9.6. Percentage of Recruits Who Assigned WSI Constructs a Given Rank

	Rank				
Construct	1st-3rd	4th-6th	7th-10th	11th-13th	14th-16th
Achievement/Effort	36.3	17.9	23.7	11.6	10.5
Adaptability/Flexibility	20.2	21.0	28.4	17.3	13.2
Attention to Detail	23.0	22.9	24.1	18.4	11.6
Concern for Others	17.0	17.1	19.7	19.7	26.5
Cooperation	15.6	22.7	25.6	17.8	18.4
Dependability	14.9	26.2	30.3	17.1	11.4
Energy	24.0	21.1	25.7	15.2	14.0
Independence	17.8	15.4	22.2	15.7	28.9
Initiative	7.5	15.9	34.8	26.8	15.1
Innovation	25.7	14.0	27.8	17.8	14.8
Leadership Orientation	23.8	21.7	24.0	18.1	12.4
Persistence	11.1	13.2	23.5	31.4	20.8
Self-Control	13.3	20.5	26.3	22.5	17.3
Social Orientation	22.4	17.9	23.0	17.6	19.0
Stress Tolerance	7.8	10.5	19.0	19.0	43.7
Cultural Tolerance	19.7	22.1	21.9	13.8	22.5

Note. $n = 630$.

Subgroup Differences

Tables 9.7 and 9.8 show mean WSI construct scores by sex and race/ethnicity, respectively. Statistically significant gender differences were found for only 4 of the 16 WSI constructs. On average, females ranked Concern for Others higher than males ($d = 0.53$). Conversely, males ranked Stress Tolerance ($d = -0.26$), Persistence ($d = -0.25$), and Self-Control ($d = -0.22$) higher than females. With regard to race/ethnicity, statistically significant differences were found for five WSI constructs. On average, Blacks ranked Achievement/Effort ($d = 0.26$), Concern for Others ($d = 0.24$), and Cooperation ($d = 0.23$) higher than did Whites. Conversely, White Non-Hispanics ranked Initiative ($d = -0.27$) higher but Cultural Tolerance ($d = 0.37$) lower than did Hispanics. Despite the statistical significance of these differences, their magnitude is small based on common effect size conventions (e.g., Cohen, 1992).

Table 9.7. WSI Scores by Gender

		Male		Female	
Construct	d_{FM}	*M*	*SD*	*M*	*SD*
Achievement/Effort	0.15	10.25	4.65	10.96	4.45
Adaptability/Flexibility	0.14	8.86	4.40	9.50	4.13
Attention to Detail	0.03	9.31	4.40	9.46	4.40
Concern for Others	**0.53**	7.05	4.68	9.51	4.75
Cooperation	0.14	8.27	4.47	8.89	4.27
Dependability	0.09	9.00	4.04	9.35	4.02
Energy	-0.16	9.66	4.50	8.95	4.51
Independence	-0.16	8.03	5.00	7.26	5.08
Initiative	0.11	7.36	3.73	7.76	3.79
Innovation	-0.06	9.16	4.67	8.86	4.75
Leadership Orientation	-0.10	9.57	4.49	9.11	4.55
Persistence	**-0.25**	7.54	4.21	6.48	4.08
Self-Control	**-0.22**	8.45	4.27	7.52	4.25
Social Orientation	-0.10	8.94	4.70	8.48	4.54
Stress Tolerance	**-0.26**	6.07	4.58	4.89	4.03
Cultural Tolerance	0.11	8.49	4.91	9.03	4.70

Note. $n_{Male} = 437$, $n_{Female} = 189$. d_{FM} = Effect size for Female-Male mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., Males) are listed second in the effect size subscript. Statistically significant effect sizes are bolded ($p < .05$, two-tailed).

Table 9.8. WSI Scores by Race/Ethnic Group

			White		Black		White Non-Hispanic		Hispanic	
Construct	d_{BW}	d_{HN}	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Achievement/Effort	**0.26**	-0.01	10.16	4.61	11.35	4.52	10.22	4.63	10.18	4.38
Adaptability/Flexibility	0.05	-0.07	9.00	4.37	9.21	4.40	9.07	4.38	8.76	4.24
Attention to Detail	0.07	-0.19	9.35	4.38	9.66	4.55	9.43	4.40	8.58	4.15
Concern for Others	**0.23**	-0.01	7.49	4.76	8.59	4.65	7.47	4.75	7.45	4.94
Cooperation	**0.24**	0.18	8.15	4.37	9.21	4.40	8.08	4.36	8.86	4.31
Dependability	0.04	-0.18	9.16	4.08	9.31	3.84	9.27	4.09	8.54	4.06
Energy	-0.19	0.06	9.47	4.51	8.62	4.54	9.52	4.51	9.80	4.38
Independence	0.01	-0.17	7.83	5.09	7.86	5.05	7.95	5.15	7.08	4.75
Initiative	-0.11	**-0.27**	7.67	3.75	7.24	3.86	7.81	3.76	6.78	3.70
Innovation	-0.11	-0.16	9.16	4.58	8.65	4.86	9.27	4.60	8.52	4.62
Leadership Orientation	-0.10	0.14	9.62	4.46	9.17	4.62	9.54	4.44	10.14	4.65
Persistence	-0.17	-0.02	7.27	4.20	6.57	4.28	7.23	4.10	7.16	4.44
Self-Control	-0.08	0.12	8.31	4.39	7.95	3.81	8.17	4.34	8.70	4.35
Social Orientation	-0.12	0.06	8.98	4.71	8.39	4.56	8.92	4.72	9.20	4.80
Stress Tolerance	-0.17	0.08	6.00	4.62	5.20	4.13	5.91	4.51	6.29	5.00
Cultural Tolerance	0.13	**0.37**	8.38	4.92	9.01	4.67	8.15	4.92	9.96	4.70

Note. $n_{White} = 413$. $n_{Black} = 94$. $n_{White\ Non\text{-}Hispanic} = 357$. $n_{Hispanic} = 83$. d_{BW} = Effect size for Black-White mean difference. d_{HN} = Effect size for Hispanic-White Non-Hispanic mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., White) are listed second in the effect size subscript. Statistically significant effect sizes are bolded ($p < .05$, two-tailed).

Correlations Among WSI Constructs

Table 9.9 displays correlations among the WSI construct scores. On average, the WSI constructs showed low intercorrelations (mean $r = -.07$). Perhaps the most striking feature of Table 9.9 is the high number of negative correlations. Nevertheless, the large number of negative correlations present among the construct scores is not surprising given the ipsative nature of the construct scores (Hicks, 1970).[42] The construct scores exhibiting the highest positive correlations tended to be those that were conceptually related. For example, correlations among Attention to Detail, Achievement/Effort, and Dependability (all related to Conscientiousness) ranged from .11 to .23, and correlations among Concern for Others, Cooperation, Social Orientation, and Cultural Tolerance (all related to Agreeableness) ranged from .06 to .33. The constructs that exhibited the largest negative differences suggested a task-oriented *vs.* person-oriented interpretation of the data. For example, recruits who tended to score high on Achievement/Effort and Attention to Detail, tended to score low on Cultural Tolerance, Concern for Others, and Social Orientation.

[42] The fully ipsative nature of the data obviated the performance of a factor analysis.

Table 9.9. Correlations Among the WSI Construct Scores

WSI Construct	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1 Achievement/Effort	-														
2 Adaptability/Flexibility	**.08**	-													
3 Attention to Detail	**.23**	.00	-												
4 Concern for Others	**-.22**	**.07**	**-.13**	-											
5 Cooperation	**-.14**	**.09**	**-.18**	**.33**	-										
6 Dependability	**.15**	**-.10**	**.11**	**-.16**	.02	-									
7 Energy	-.01	**-.18**	-.01	**-.19**	**-.15**	.03	-								
8 Independence	**-.16**	-.03	**-.07**	-.06	**-.12**	**-.07**	-.05	-							
9 Initiative	.02	-.04	-.04	**-.13**	**-.14**	-.01	**-.08**	.01	-						
10 Innovation	**-.16**	**-.09**	**-.14**	-.01	**-.11**	**-.16**	**-.09**	.06	-.03	-					
11 Leadership Orientation	.00	**-.25**	**-.09**	**-.18**	**-.20**	-.04	-.02	-.05	-.02	.03	-				
12 Persistence	.05	**-.15**	.05	**-.24**	**-.21**	-.01	**-.10**	.01	.06	-.02	.03	-			
13 Self-Control	**-.22**	**-.17**	**-.10**	**-.18**	**-.18**	**-.16**	.03	**-.08**	**-.10**	**-.08**	-.03	.04	-		
14 Social Orientation	**-.18**	-.01	**-.23**	.06	**.07**	**-.14**	**-.07**	**-.32**	**-.10**	**-.09**	**-.07**	**-.19**	.03	-	
15 Stress Tolerance	**-.10**	**-.15**	**-.08**	**-.26**	**-.18**	-.04	**.07**	**-.14**	-.03	**-.18**	-.03	.02	**.27**	.01	-
16 Cultural Tolerance	**-.28**	-.03	**-.24**	**.19**	**.07**	**-.25**	**-.17**	**-.08**	**-.16**	.00	**-.07**	**-.18**	-.06	**.16**	-.06

Note. n = 630. Correlations are Spearman rank-order correlations. Bolded correlations are statistically significant ($p < .05$, one-tailed).

Discussion

The WSI provides an innovative means for assessing personality. The measure can be labeled as a "POP-Hybrid" in that it supports a variety of scoring options that permit its use as a measure of person-organization (PO) fit or as a measure of personality (P). The WSI incorporates several strategies to mitigate response distortion: (a) a forced-choice rank-order format, (b) the use of foil (i.e., unscored) constructs, (c) an instruction set that directs respondents to consider work they might perform well rather than describing their personality per se, and (d) the possibility of supporting multiple criterion-specific scoring algorithms. Although one could still provide optimal responses to the WSI for a given criterion, the goal is to obtain sufficiently different scoring algorithms that a respondent cannot hope to attain high standing on all criteria with a single optimal rank-order profile.

The analyses conducted to date provide a limited glimpse into the likely success or failure of the instrument. Data collected during the pilot test, faking research, and field test phases of the project all provide promising results for those limited areas where we could examine its performance.

- Convergent validity results of the WSI with a marker test of the Big Five personality constructs were promising.
- Analysis of subgroup differences showed some rather small differences between focus and reference groups.
- Respondents did alter their rankings between the honest and faking conditions, but other evidence was consistent with a resistance to faking: (a) the WSI still differentiated among recruits in the faking conditions; (b) correlations between honest and faked WSI construct

scores were generally low, indicating notable individual differences in either recruits' ability to fake the WSI effectively and/or their compliance with the faking instructions; and (c) the mean correlation between recruits' profiles and the target Army profile (which increased between the honest and faking conditions) remained low in the faking conditions—all of which are consistent with a resistance to faking. Despite this promise, the only certain way to determine susceptibility to faking is to administer the WSI in an operational setting.

The primary evaluation of the WSI during this project will occur in the concurrent validation. At that time, we will have criterion data against which various WSI scoring algorithms may be examined, which will permit us to determine the correlations of WSI scores with the various target criteria. We will also begin to understand the degree to which different optimal scoring algorithms are realized.

Next Steps for the WSI

No modifications are proposed for the content or structure of the WSI for the concurrent validation effort. Data from the WSI that were gathered from recruits during the pilot, faking research, and field test data collections will be included in the Select21 attrition database. As this database matures, we will examine relations between WSI data and attrition at various stages of the initial enlistment term. As part of this analysis, we will begin to explore alternative scoring algorithms for the WSI that attempt to maximize its relation to our various criterion measures. Assuming promising results in the concurrent validation, the next step would be to evaluate the WSI under operational conditions.

At some point, we would also like to gather test-retest data on the WSI to assess (a) the consistency of individuals' construct rankings across occasions, and (b) the average consistency of construct rankings across occasions (discussed further in Chapter 15).

CHAPTER 10: PREDICTOR SITUATIONAL JUDGMENT TEST (PSJT)

Gordon W. Waugh and Teresa L. Russell
HumRRO

Background

Prior research suggests that situational judgment test (SJT) scores are likely to predict supervisor's ratings of job performance (McDaniel, Finnegan, Morgeson, Campion, & Braverman, 2001) and provide incremental validity over the ASVAB (Knapp et al., 2002; Peterson et al., 1993). With this in mind, we developed a predictor SJT—the PSJT. In this chapter, we provide a short overview of the development of the PSJT and describe the field test results in some detail.

Overview of PSJT Approach

The purpose of the PSJT is to *predict* first-tour job performance in the Future Force by simulating situations a Soldier faces prior to enlistment. While the general process of developing the PSJT is similar to that of the Criterion Situational Judgment Test (CSJT; described in Chapter 5)—generate scenarios, generate actions, develop scoring key—the two instruments differ in content and possible scoring mechanisms.

Content

Clearly, the CSJT should contain military scenarios. The case is not so clear, however, for the PSJT. On the one hand, an instrument with some *military* scenarios might have greater face validity than one that contains *civilian* scenarios. On the other hand, if military scenarios were deeply steeped in the military setting, they might require tacit knowledge of the Army that Army applicants could not be expected to have. Our strategy was to construct a pilot test version of the PSJT containing both civilian and military scenarios.

The PSJT could have been targeted to measure either the Select21 performance dimensions or the Select21 knowledges, skills, and attributes (KSAs). We decided to use the performance dimensions because the behaviorally worded performance dimension definitions would be more useful than the more trait-like KSAs when generating critical incident and scenarios. In contrast to the CSJT, however, scenarios focused on experiences very early in a Soldier's first enlistment term.

Because the PSJT items ask respondents what *should* be done in a situation, the PSJT taps tacit knowledge and judgment rather than motivation or skill. Accordingly, we selected the following six performance dimensions (from the Select21 job analysis) that we judged could be assessed using an SJT format: Exhibiting Effort and Initiative, Adaptability to Changing Conditions, Relating to and Supporting Peers, Effective Self-Management, Effective Self-Directed Learning, and Teamwork. These performance dimensions are defined in Figure 5.1 of Chapter 5. Although Exhibiting Effort and Initiative taps motivation, we included it because we thought that some scenarios written to target this dimension might also tap tacit knowledge.

Scoring Schemes

Our plan was to develop and compare three possible scoring schemes for the PSJT: the traditional approach, a personality-based approach, and a style-based approach. The traditional approach yields a single total score; applicant's responses are compared to the key and summed across items. We refer to this as the "judgment key." The personality-based approach pioneered by Steve Motowidlo and his colleagues yields scores for selected personality traits (Motowidlo, Diesch, & Jackson, 2003). The style-based approach was used experimentally in the Army's NCO21 project (Knapp et al., 2002). It yields scores for different response styles such as "take the easy way out" or "express concern." To conserve project resources, we postponed development of the style-based approach until the concurrent validation.

While we planned to compare these scoring schemes, it is important to note that our development approach for the PSJT was a traditional one. That is, we generated scenarios and response options using subject matter experts (SMEs); our staff provided input and edited the items. We chose the traditional approach because it has resulted in good zero-order correlations and incremental validity in the past (McDaniel et al., 2001; Peterson et al., 1993). We did not deliberately construct the PSJT to be a personality measure—in which case we would probably have asked psychologists to write most of the items in a particular manner—nor did we dictate a style-based instrument. We developed different schemes for scoring an SJT, each approach using the same test items.

PSJT Development

The PSJT development phase involved generating and selecting military scenarios, generating and selecting civilian scenarios, writing response options, and developing judgment and personality-based scoring keys.

Generate and Select Military Scenarios

Two principles defined the approach to generating military scenarios. First, the PSJT should include early-career scenarios that would not require much tacit knowledge of Army life. Therefore, we planned to collect scenarios from drill sergeants and Advanced Individual Training (AIT) and One Station Unit Training (OSUT) instructors at Forts Jackson, Leonard Wood, Lewis, Eustis, Benning, and Gordon[43]. Second, the PSJT should not necessarily be restricted to performance dimensions from the job analysis. We wanted to identify naturally occurring critical incident categories for early career first-term performance to ensure that we covered important early career dimensions. Toward that end, we began the data collection with several critical incident generation workshops and workshops where participants wrote critical incidents first and then wrote scenarios targeted toward six performance dimensions from the job analysis.

A total of 45 instructors participated in the workshops. They wrote approximately 300 unconstrained incidents. Workshop participants also wrote scenarios based on situations new Soldiers encounter during Basic or AIT/OSUT. Three staff members independently sorted the critical incidents, printed on cards, to identify naturally occurring categories. The staff members

[43] During the first four workshops (i.e., Fort Jackson and Fort Leonard Wood), drill sergeants and AIT/OSUT instructors wrote both critical incidents and scenarios.

met to reach consensus on the categories, wrote definitions for the final categories, and categorized incidents into the final categorization scheme. The final consensus-based categories appear in Table 10.1.

Table 10.1. Basic/AIT/OSUT Critical Incident Dimensions

A. Teamwork	Understands own and team tasks in relation to the mission or assignment; coordinates and helps team members to ensure the team will achieve its goals. Versus: not performing own tasks so that others have to do them.
B. Support for Peers	Attends to and supports other team members; notices aberrant and potentially self-destructive behaviors of others (e.g., withdrawal, not eating); notices when others are ill, injured, or distressed; offers assistance; informs the NCOIC of problems. Versus: Ignores problem behaviors or does not realize that those behaviors are dangerous; harasses the other Soldiers; fails to report problems to the NCOIC.
C. Peer Leadership	Gives direction; leads peers when given a leadership role; gives clear instructions; distributes tasks; attempts to gain others' cooperation; devises ways to address team deficiencies; obtains the assistance of the DI as appropriate in dealing with disrespectful behavior; disciplines when appropriate. Versus: Does not give direction; fails to lead when required; gives unclear instructions; alienates subordinates; fails to discipline.
D. Safety Consciousness	Foresees and alerts others to potential hazards; follows proper safety procedures; handles emergencies in a calm, task-oriented manner. Versus: Acts without thinking about safety implications; disregards proper procedures; skips steps or takes shortcuts on tasks; falls asleep while guarding fires; uses weapons, vehicles, tools, or equipment recklessly; fails to inform others of potential hazards.
E. Respect for Authority and Orders	Follows superior's orders willingly; talks in a respectful manner to superiors; informs NCOIC of violations of regulations or orders. Versus: Ignores or refuses to follow orders; complains about orders; talks back or speaks disrespectfully to superiors; fails to report violations of regulations or orders.
F. Delinquency	Resists temptation to steal military property; informs NCOIC of thefts. Versus: Steals food, property, tobacco, or other articles; lies to cover up own behavior.
G. Self-Control	Resists temptation to indulge in prohibited activities such as fraternization, and tobacco, food and drug use; follows Army regulations. Versus: Fails to control own behavior; hides and indulges in contraband items such as food and tobacco; is caught in acts of sexual misconduct; loses control of temper; gets into fights; violates Army regulations.
H. Self-Management	Keeps self and personal gear and equipment in a ready state; seeks treatment and assistance if sick or injured; has needed items on hand; maintains control of personal weapons. Versus: Attempts to work while sick or injured, may become a casualty; consumes unsafe water or food; loses weapon or gear; forgets needed items.
I. Physical Fitness and Endurance	Takes initiative to develop own physical fitness, seeking assistance from peers and supervisors as needed; loses weight and becomes fit; endures difficult weather and other obstacles to accomplish physical tasks. Versus: Avoids exercise, shirks physical training; gives up easily.
J. Adaptability	Accepts changes with a positive attitude; improves own behavior when counseled. Versus: Allows homesickness for loved ones or emotional problems to interfere with training; withdraws from peers; is not receptive to attempts to communicate; exhibits dangerous behavior; attempts suicide.
K. Motivation, Effort, and Initiative	Volunteers for assignments; performs duties, even under adverse conditions; prepares for assignments; performs duty to a high degree of proficiency; spends beyond the required amount of time or effort. Versus: Doesn't try; exhibits minimal effort; plays hooky; malingers.

As previously mentioned, we had selected six performance dimensions from the Select21 job analysis that we thought could be assessed using an SJT format. After the first few workshops, we examined the content of the generated scenarios. Two performance dimensions, Teamwork and Support for Peers, were yielding more potentially useful scenarios than the others. Even so, there were some scenarios that we could make into items for most of the other selected performance dimensions. We were, however, concerned about the scenarios for Exhibiting Effort and Initiative. For those scenarios, the only appropriate course of action was too obvious. Therefore, we dropped that performance dimension from the test plan for the PSJT. The final five performance dimensions included on the PSJT were Adaptability to Changing Conditions, Relating to and Supporting Peers, Effective Self-Management, Effective Self-Directed Learning, and Teamwork.

Next, we sorted the scenarios into the critical incident categories to see if any unique or different constructs were represented. Based on that analysis, we developed two recommendations. First, a number of incidents from Basic Combat Training (BCT) involved a dilemma in which the Soldier had to decide whether to participate in inappropriate behavior with peers. The "correct" response option (according to instructors) is to report the incident to the instructors. We are unsure how this situation would play out in the civilian world and how another set of raters might evaluate potential behaviors. We decided to include items of this type on the PSJT and assess how well they worked. Second, we received a number of scenarios relevant to the Peer Leadership critical incident category. Within the Select21 performance dimension taxonomy, these peer leadership scenarios were usually categorized into "Teamwork." They were low-level initiating structure behaviors. We included scenarios for both teamwork and peer leadership on the SJTs with the hope they would contribute to the predictive validity of the measure.

All of the scenarios were typed into a relational database. We recorded several attributes of each scenario, such as its target dimension, relevance to the future military, relevance to the civilian sector, status in the data collection efforts, and so on. In turn, staff assessed the potential usefulness of each scenario obtained in the workshops. When evaluating each scenario, we considered the following characteristics of a well-designed scenario:

- There are several possible response options (i.e., actions) for the scenario.
- There are several possible response options that some people will choose as best.
- The potential response options differ in effectiveness.
- The scenario is likely to be relevant in the future.
- The scenario is relevant to Soldiers during Basic or AIT/OSUT. (A staff member with considerable knowledge in this area made this judgment.)
- The respondent has all the information needed to answer the question.
- The wording is clear and succinct.

If a scenario did not possess these characteristics, we tried to improve it. If we were unsuccessful, we dropped it.

Generate and Select Civilian Setting Scenarios

We planned to include approximately 50 civilian scenarios in the PSJT. Toward that end, we generated civilian setting scenarios with (a) college students at George Mason University and (b) AIT/OSUT students who, having recently been civilians, were able to draw on appropriate experiences.

We gave AIT/OSUT students at Forts Knox, Eustis, Gordon, and Huachuca military scenarios (written in prior workshops with instructors) and asked them to write civilian-setting scenarios that were similar to those early-career military scenarios. In writing their scenarios, Soldiers were asked to think about their experiences at work, at school, and in extracurricular and athletic activities that are parallel to the military scenarios. They were to avoid topics that might be too sensitive or narrow for test items, such as scenarios based on religious activities, cultural events, or gender issues. After students had been writing scenarios for 10 minutes or so, we stopped the group and asked volunteers to read their scenarios out loud. We then circulated through the group and read scenarios as they were being written. This allowed us to make on-the-spot corrections and edits. We had anticipated that writing civilian scenarios might be difficult, but the students seemed to have no trouble doing so.

We also collected civilian setting scenarios from George Mason University students. In the first part of an hour-long workshop, we presented military scenarios and asked the students to write direct civilian counterparts to those scenarios. In the second part of the workshop, we solicited civilian scenarios for the performance dimensions that were of interest to us, without regard to whether the scenario had a direct military counterpart.

Four hundred and fifty-two scenarios were generated. We retained only the best parallel civilian scenario for each of the associated 62 military scenarios. The "best parallel" scenarios were ones that (a) retained the key features and the problem presented in the original scenario and (b) provided a relatively commonplace civilian scenario that most people would understand. This screening process resulted in 62 civilian scenarios. Eight additional usable scenarios were retained because they were not similar to any of the 62 selected scenarios. Thus, a total of 70 civilian scenarios were retained at this stage.

Write Response Options and Collect Preliminary Effectiveness Ratings

Once the scenarios were written, we asked E5–E7 NCOs at Forts Benning, Gordon, Knox, and Eustis to describe actions that a new Soldier might take in each scenario. The vast majority of these NCOs were drill sergeants and AIT instructors. We instructed participants to imagine that they are the person in the situation and to think of at least three response options. To help participants get started writing response options, we gave them tips for thinking of actions that might make good response options. Staff members reviewed, condensed, and edited response options with the goal of having 6–9 response non-redundant, usable options for each scenario.

To help evaluate and edit the revised response options, we asked some of the NCOs to rate the effectiveness of the response options. Using their data, we identified response options

with low standard deviations on the effectiveness ratings and, for each scenario, a set of response options representing a range of effectiveness.

Develop Personality-Based Scoring Scheme

We conducted an expert judgment exercise with HumRRO and ARI research staff to develop the personality-based scoring scheme. We started by identifying the Select21 KSAs that linked to the five performance dimensions targeted by the PSJT. During the job analysis, 17 experts (personnel researchers from HumRRO and ARI) linked the Select21 KSAs to the performance dimensions; 17 KSAs were linked to one or more of the five performance dimensions. Those 17 were our initial candidates for the exercise. Four of the linked KSAs were cognitive abilities and two were skills. Since our focus was on a personality-based scheme, we dropped those six from further consideration. We dropped another three KSAs because we believed that they would not be manifested in the scenarios or would be too difficult to link to behaviors in scenarios. The remaining eight KSAs were included in the exercise:

- Achievement Orientation
- Self-Reliance
- Dependability
- Affiliation/Sociability
- Agreeableness
- Social Perceptiveness
- Team Orientation
- Intellectance

In the exercise, the experts judged the strength of the relationship between the examinee's standing on a particular trait and his/her rating of the effectiveness of each response option. We told the experts to think of this as a correlation between the scores on a trait and the effectiveness ratings likely to be given to the response options. Specifically, they rated the strength of the relationship between the examinee's standing on each of the traits and his/her rating of the effectiveness of the response option using the scale shown on the next page. A range of correlations was associated with each scale point. The experts were told to consider the traits to be perfectly measured, corrected for unreliability.

–4	Very high negative relationship (r = –.50 to –1.00)
–3	High negative relationship (r = –.30 to –.49)
–2	Moderate negative relationship (r = –.15 to –.29)
–1	Low negative relationship (r = –.05 to –.14)
0	No relationship (r = –.04 to .04)
1	Low positive relationship (r = .05 to .14)
2	Moderate positive relationship (r = .15 to .29)
3	High positive relationship (r = .30 to .49)
4	Very high positive relationship (r = .50 to 1.00)

At this point, the PSJT contained 63 civilian scenarios and 50 military ones, almost all with five to seven response options. We constructed two forms, each containing half of the

current PSJT scenarios. Each form had two parts: a military part and a civilian part. Sixteen experts completed their assignments, and one rater completed half of the assignment. A few raters completed both the military and civilian portions of their forms. When assigning raters to forms, an attempt was made to balance the assignments so that each form had an equal number of HumRRO and ARI raters. The assignments were also somewhat balanced in terms of the raters' familiarity with the Army. Each form had five or more raters.

To assess the consistency with which raters made their judgments, we computed interrater reliability estimates by form. The single-rater estimate ICC(C,1) and adjusted five raters ICC(C,5) appear in Table 10.2. The single-rater reliabilities make comparisons across forms simple. The five-rater reliability is important because it is the best estimate of the reliability of the mean ratings that we planned to use for traitedness scoring. As shown, the reliability estimates for all of the traits except Intellectance were reasonably high. The mean ICC(C,5) for Intellectance was only .51, but it ranged from .74 to .84 for the other seven traits. We dropped Intellectance from the list of PSJT-relevant traits.

Table 10.2. Interrater Reliability Estimates for Traitedness Ratings

	Form ICC(C,1)				
Trait/Metric	Civilian 1	Civilian 2	Military 1	Military 2	Mean
Achievement Orientation	.54	.43	.42	.40	.45
Self-Reliance	.49	.37	.41	.35	.40
Dependability	.53	.40	.45	.57	.49
Sociability	.47	.37	.42	.35	.40
Agreeableness	.61	.39	.42	.25	.42
Social Perceptiveness	.50	.29	.39	.31	.37
Team Orientation	.57	.44	.55	.52	.52
Intellectance	.24	.13	.27	.10	.19
Mean of column	.49	.35	.42	.36	.41

	Form ICC(C,5)				
Trait/Metric	Civilian 1	Civilian 2	Military 1	Military 2	Mean
Achievement Orientation	.85	.79	.78	.77	.80
Self-Reliance	.83	.75	.77	.73	.77
Dependability	.85	.77	.80	.87	.82
Sociability	.82	.75	.79	.73	.77
Agreeableness	.89	.76	.79	.62	.76
Social Perceptiveness	.84	.67	.76	.69	.74
Team Orientation	.87	.80	.86	.84	.84
Intellectance	.61	.42	.65	.37	.51
Mean of column	.82	.71	.78	.70	.75

We used the traitedness judgments to create a key for the PSJT. During the field test (described later in this chapter), we tried out different methods of using the PSJT data to create the key.

Data Collection

Pilot Testing

The primary purpose of the pilot test was to investigate and resolve three primary issues: (a) Can we create a reasonably reliable test form composed of only civilian scenarios? (b) Can we overcome central tendency concerns regarding the effectiveness key? and (c) Is the personality-based key reasonably reliable?

Sample

Four hundred and twenty new recruits at the Fort Jackson, Knox, and Benning reception battalions participated in the PSJT pilot testing. After reviewing administrator logs and recruit data, we determined that three recruits may not have responded conscientiously and omitted them from further analysis, making our final sample size 417. Each recruit completed one of four civilian PSJT forms (A–D) and one of four military PSJT forms (1–4). There were four pairings of forms: A-1, B-2, C-3, and D-4. Within each form-pair, the order was randomized. That is, half of the recruits got the military form first; the other half got the civilian form first. Most items had seven response options. The civilian forms had 14–16 items; the military forms had 11–13 items. There was no attempt to put a military item and its parallel civilian item within the same form-pair.

Initial Calculation of Judgment Scores

The recruits responded by rating the effectiveness of each option on a 7-point scale (where higher numbers represent greater effectiveness). We computed the judgment score for each response option using Equation 1 below.

$$\textit{Judgment Score}_{\textit{Option } x} = 6 - |\, \text{Recruits}\textit{Rating}_{\textit{Option } x} - \textit{keyedEffectiveness}_{\textit{Option } x} \,| \qquad (1)$$

The keyed effectiveness ratings were based on data collected earlier from the AIT/OSUT instructors.

We subtracted the difference between the rating and keyed effectiveness values from 6 to reflect the scores (i.e., so that higher values would represent better scores).[44] The judgment score for an entire test form was the mean of the option scores.

The reliability estimates of the judgment scores for both the military and civilian forms were reasonably high. As shown in Table 10.3, the reliability estimates are around .90. The reasonably high correlations between military and civilian form pairs (r = .70 to .85, and r_c = .83 to .95) suggest that the judgment score is measuring essentially the same thing on the civilian and military forms. The correlations between these forms are almost as high as the reliability estimates.

[44] We planned to collect data from ANCOC students to establish the keyed effectiveness values, but that data collection had not yet occurred. In the interim, we used the effectiveness ratings made by E5-E7 NCOs at Forts Eustis, Gordon, Knox, and Benning as the keyed effectiveness values.

Table 10.3. Correlations between Civilian and Military Forms and Reliability Estimates for Judgment Scores

Form Pair			Coefficient Alpha	
(Civilian, Military)	r	r_c	Civilian	Military
A, 1	.76	.83	.92	.92
B, 2	.85	.95	.91	.88
C, 3	.70	.83	.87	.83
D, 4	.82	.90	.91	.90

Note. $n = 79$. Each soldier completed only one form pair. r_c is corrected for attenuation due to unreliability. All correlations are significant at $p < .0001$.

These data suggested that a form including only civilian scenarios (which are likely to be more relevant to the applicant population) would be about as reliable as the military form and would measure essentially the same constructs. This finding, together with our concern that military scenarios could be inappropriate for selection testing of applicants, led us to conclude that it would be reasonable to include only civilian scenarios on the field test version of the PSJT.

Judgment Scoring Key Adjustments

An effectiveness rating-based scoring key has a potential disadvantage. The variability of an examinee's responses is highly correlated (in a negative direction) with the judgment scores. The key has a ceiling and a floor because it is the average of the SMEs' effectiveness ratings. That is, an item rarely has a keyed score of "1" or "7" because there is a central tendency effect. In turn, the central tendency effect makes two relatively simple coaching strategies possible. An examinee could get a fairly good score by simply rating every option a 4 (the middle of the rating scale) or by avoiding using ratings of "1" or "7" (Cullen, Sackett, & Lievens, 2004).

We investigated three methods of mitigating the potential coaching effects—truncating the scores, stretching the key, and rank ordering the scores. To truncate the scores, we simply converted responses of "1" to "2" and "7" to "6" and recalculated the test scores. To stretch the scoring key, we used the following formula:

For original key values above 4.0, newValue = oldValue + 0.5 * (oldValue – 4).
For original key values below 4.0, newValue = oldValue – 0.5 * (4 – oldValue).

For rank ordering, we rank-ordered the effectiveness scores to form a new key. The rank score for an option was the difference between the keyed rank and the rank assigned by the Soldier. This difference was always expressed as a positive number.

For each scoring method, we computed coefficient alpha, the correlation between the total score and the standard deviation of the ratings of a hypothetical respondent who gave all perfect responses, and the effect size—the standardized difference between the test score mean in our data and the test score computed using the coaching method. As shown in Table 10.4, the coaching methods have a substantial effect on scores when no scoring adjustments are made (i.e.,

d = .73 and .69). Truncating the scores resulted in even worse effects. Stretching and ranking resulted in minimal coaching effects with acceptable estimated reliabilities.

Table 10.4. Comparison of Key Adjustment Methods

Adjustment	Coefficient Alpha	Correlation Between Key and *SD* of Ratings	Effect Size for Coaching Method 1 "4"	No "1" or "7"
Unadjusted	.84	–.54	.73	.69
Truncated	.84	–.57	.85	.81
Stretched	.76	–.22	–.50	.27
Ranked	.77	.06	.00	.00

Note. All statistics are averaged across eight forms. The number of options per form ranges from 28–56 with an average of 44.

Personality-Based Scores

To investigate the potential usefulness of the personality-based scoring key, we computed the reliability estimates for the seven trait scales on each form and averaged those estimates across forms. As shown in Table 10.5, the alphas for all of the traits, except one, Social Perceptiveness, were acceptable. However, the number of response options scored on each trait is highly variable and the range of reliability estimates is partly a consequence of varying numbers of options for traits. The alpha adjusted to 20 items estimates what alpha would have been had each trait been scored on 20 options. As shown, these alphas are less variable and are still acceptable.

Table 10.5. Average Trait Scale Reliability Estimates Computed Across Eight Forms

Trait Scale	Alpha	Alpha (k=20)
1. Achievement Orientation	.74	.68
2. Self-Reliance	.71	.67
3. Dependability	.75	.69
4. Sociability	.71	.76
5. Agreeableness	.67	.63
6. Social Perceptiveness	.49	.57
7. Team Orientation	.70	.67

We were, at this point, considering using a rank-order rather than effectiveness ratings response format in the field test version of the PSJT. With that in mind, we also computed a version of the personality scores using rank-ordering. We found that most of the alphas based on ranks were negative, possibly due to the ipsative nature of the data. Thus, we favored the effectiveness rating scale over a rank-ordered rating scale.

Conclusions

The pilot test data led us to conclude that it would be reasonable to develop a test form using only civilian scenarios using an effectiveness rating scale. Data also suggested that the central tendency effect on the scoring key could be reduced using stretching or ranking, but ranking created other problems for the personality keys. Thus, we chose stretching as the

preferred method. Finally, the results suggested that the personality keys were reliable enough to retain for further investigation in the field test.

Faking Research

Sample and Study Design

We developed a shortened PSJT (PSJT-S) for the faking research effort. It had 12 items; each item had seven response options (i.e., a total of 84 response options). Because the PSJT asks respondents what they "should" do—rather than what they "would" do—faking is not an issue with the SJT. However, it is possible that coaching might improve scores. Therefore, the effects of coaching were examined.

The PSJT-S was administered twice to 199 new recruits, once in an "uncoached" condition and once in a "coached" condition. In the uncoached condition, the recruits were asked to pretend that they were applying to join the Army, that they wanted to be accepted, and that their scores would affect the acceptance decision. It is likely that scores in the uncoached condition would be similar to scores obtained in the pilot test and field test (in which no motivational instructions were given). Participants in the coached condition were given a written list of tips: which type of options to give higher ratings to and which type of options to give lower options to. The script for the two conditions is provided in Appendix F. The uncoached condition always occurred first. Had coaching occurred first, it might have influenced the results of the uncoached condition.

Results

After coaching, the recruits tended to give higher effectiveness ratings. On average, the ratings went up by .24 scale points, which is .48 standard deviations. Even so, the judgment scores (see Formula 1) went down significantly ($t = 9.15$, $df = 197$, $p < .0001$) by .67 standard deviations (using the honest group standard deviation) after coaching. The mean total score dropped from 4.44 to 4.22 ($SD_{\text{uncoached}} = .33$, $SD_{\text{coached}} = .33$, $SD_{\text{difference}} = .35$). Among the 84 response options, scores went up significantly after coaching for only 3 options; scores went down significantly for 57 options.

The internal consistency reliability estimates (coefficient alpha) for the two test administrations were high and comparable (.88 for the uncoached condition and .92 for the coached condition). The total scores for the two conditions correlated .59. Because coaching did not improve the scores, no changes were made to the PSJT based on the faking research results.

Final Scoring Key Development

Sample and Study Design

We had developed a preliminary scoring key to use in PSJT development and pilot testing using data collected from AIT/OSUT instructors. Since the PSJT needed to be ready for the field test before the final scoring key was ready, we used the preliminary key to make decisions concerning which options and items to use for the field test.

Subsequently, we collected effectiveness-ratings data from 71 ANCOC students and 4 AIT instructors. These data were collected at Fort Lee, Fort Knox, and the Aberdeen Proving Ground to serve as the scoring key for the PSJT in the field test and concurrent validation. We divided the PSJT items across two forms and asked each NCO to complete one test form. NCOs who finished early were asked to complete the second form. We dropped eight NCOs because they either had many missing responses or had very low correlations with the other NCOs. The final key was based on these 67 NCOs. The number of NCOs rating an item ranged from 55 to 67.

Results

Because there were many missing values, a traditional computation of interrater reliability would require the listwise deletion of all NCOs with any missing values. In contrast, Rasch analysis (i.e., Item Response Theory 1-paramater logistic model) computes internal consistency reliability estimates without excluding people with missing responses. The Rasch analysis computed a reliability estimate of .97 using the 55–67 NCOs per option. The reliability of ratings based on a single NCO was estimated at .33 using the Spearman-Brown prophecy formula.

General Description of the PSJT

Our goal was to develop, by the end of the field test, a concurrent validation version of the PSJT (PSJT-CV) that (a) had reasonably good psychometric properties and (b) could be administered in one hour or less.

The PSJT-FT (field test version) had two forms (A and B) with 32 unique items on each form. Each recruit was given 60 minutes to complete one of the two PSJT forms. Each test item had a stem—a description of a civilian scenario or situation. Each item contained up to 7 response options—actions that could be taken in the situation. Recruits rated the effectiveness of each response option (i.e., action) using the following 7-point rating scale shown in Figure 10.1.

Ineffective action.		Moderately effective action.			Very effective action.	
The action is likely to lead to a bad outcome.		The action is likely to lead to a passable or mixed outcome.			The action is likely to lead to a good outcome.	
—— Low ——			**——— Moderate ———**		**—— High ——**	
1	**2**	**3**	**4**	**5**	**6**	**7**

Figure 10.1. PSJT response option rating scale.

Field Test

The primary purpose of the field test was to develop a final version of the PSJT items and scores for use in the concurrent validation. Most analyses were targeted toward examining personality and judgment scoring keys, identifying the response options and items for the final form, and estimating the psychometric properties of the final form. We identified scores that should be retained for further analyses and computed descriptive statistics for them.

Sample

Six hundred and seventy-two new recruits participated in the PSJT field testing. We dropped 137 of these because they responded to fewer than 90% of the response options. This large number is likely due to insufficient time rather than carelessness. The test administrators had reported that several recruits were unable to finish the PSJT. After reviewing administrator logs and recruit data, we determined that 15 additional recruits may not have responded conscientiously; we omitted these recruits' data from further analysis. The final sample size was 520. Based on the logs, we also estimated that the concurrent validation version of the PSJT would need to have 26 items in order for most Soldiers to complete it in one hour.

Judgment Scores

The keyed value for each option was computed using the *stretched-key* method described previously for the pilot test—with one modification. After stretching the key, each key value was rounded to the nearest whole number. If the number was less than 1 (i.e., the minimum rating scale value), then it was changed to 1; if the number was greater than 7 (i.e., the maximum rating scale value), then it was changed to 7. As a result, all of the option judgment scores were whole numbers. Converting the key values to integers reduced the internal consistency reliability by only .005. Stretching the scoring key reduced internal consistency reliability by only .020.

The first step in the PSJT analyses was to examine the psychometric properties of the judgment scores from the two forms. As shown in Table 10.6, both forms yielded reasonably high estimates of reliability for the judgment scores.

Table 10.6. Reliability Estimates for Judgment Scores by PSJT Form

	n	M	SD	$k_{\text{response options}}$	alpha
Form A	264	4.44	.36	210	.93
Form B	256	4.54	.34	211	.93

Personality-Based Scale Scores

All of the personality key-related analyses reported in this chapter so far have used a scoring key that allows response options to be scored on multiple traits (i.e., *non-unique* weighting). For example, one response option might be scores on both Achievement Orientation and Self-Reliance. To make the personality scales as independent as possible, we selected one trait to be scored for each response option based on the traitedness judgments (*unique* weighting). Within each of these two scoring methods, we also tried using both *unit* traitedness weights and *exact* weights. In unit weighting, the recruit's rating was multiplied by one; in exact weighting, the rating was multiplied by the mean traitedness rating among the traitedness judges. Table 10.7 shows the internal consistency reliability estimates for these four trait scoring methods.

Table 10.7. Comparison of Methods of Keying Personality Scale Scores for the Full FT Form

	Coefficient Alpha				Number of Options	
	Unit Weights		Exact Weights			
Trait	Unique	Non-Unique	Unique	Non-Unique	Unique	Non-Unique
1. Achievement Orientation	.85	.89	.84	.88	40	59
	.80	.91	.79	.91	30	70
2. Self-Reliance	.52	.84	.51	.83	22	53
	.52	.85	.55	.84	25	55
3. Dependability	.74	.89	.75	.89	20	55
	.84	.92	.84	.92	30	70
4. Sociability	.56	.84	.59	.84	13	36
	.50	.87	.50	.87	12	38
5. Agreeableness	.53	.78	.58	.79	11	44
	.54	.80	.57	.81	25	54
6. Social Perceptiveness	.64	.75	.66	.78	15	34
	.35	.73	.42	.74	6	26
7. Team Orientation	.64	.81	.69	.81	29	49
	.81	.91	.81	.91	23	56

Note. n varied from 188–249. Within each cell, the Form A value is above the Form B value. The *Number of Options* column represents the number of options scored on that trait.

As Table 10.7 shows, each trait has far more scored options using the non-unique method. Consequently, reliability estimates are higher using the non-unique scoring method. Overall, reliability was roughly the same for unit weighting vs. exact weighting. Across forms and traits, non-unique weighting had a coefficient alpha of .84 (using either unit or exact weighting) whereas unique weighting had an alpha of .63 using unit and .65 using exact weighting.

Scale intercorrelations. We computed personality scale intercorrelations using all four scoring methods. The mean intercorrelations for each scoring method appear in Table 10.8. The correlations are very high for the non-unique scoring method. Although the underlying constructs are related, these correlations are much too high. It appears that the confounding nature of the non-unique scoring method artificially raised the correlations by an average of .28.

Table 10.8. Mean Personality Scale Correlations for the Full FT Form

	Mean Interscale Correlations Across Traits			
	Unit Weights		Exact Weights	
	Unique	Non-Unique	Unique	Non-Unique
Form A	.48	.77	.47	.75
Form B	.53	.81	.54	.82

Note. n varied from 188–249.

Therefore, we used one of the unique scoring methods for remaining analyses. Unit weighting was chosen because it is simpler than exact weighting, and the two methods' reliabilities and scale correlations are similar. Thus, we used the unit, unique weighting method to compute the personality scale scores for the remainder of the analyses. Table 10.9 shows the scale correlations using the unit unique weighting, and Table 10.10 shows the differences

between the Form A and Form B correlation matrices. Even using the unique weighting method, some scale intercorrelations are high.

Table 10.9. Mean Personality Scale Intercorrelations for Full FT Form

Scale	J	AO	SR	D	S	A	SP	TO
Judgment (J)		.49	.11	.67	.49	.58	.45	.66
Achievement Orientation (AO)	.56		.56	.76	.60	.39	.59	.74
Self-Reliance (SR)	.33	.63		.39	.34	.09	.32	.40
Dependability (D)	.71	.71	.47		.60	.60	.63	.81
Sociability (S)	.51	.65	.47	.60		.44	.45	.65
Agreeableness (A)	.62	.26	.09	.44	.32		.49	.62
Social Perceptiveness (SP)	.74	.42	.17	.55	.38	.65		.65
Team Orientation (TO)	.61	.70	.50	.64	.60	.43	.47	

Note. The unit-unique trait weighting scheme was used. Form A correlations are in the lower left triangle, Form B correlations are in the upper right triangle.

In general, the Form A and Form B correlation matrices are similar (see Table 10.10). This suggests that the personality scales are measuring similar things in the two forms. Social Perceptiveness and Agreeableness show the biggest differences between forms.

Table 10.10. Between-Form Differences in Personality Scale Intercorrelations for Full FT Form

Scale	J	AO	SR	D	S	A	SP	TO
Judgment (J)		.07	.21	.04	.02	.04	.29	–.05
Achievement Orientation (AO)	.07		.07	–.05	.05	–.13	–.18	–.04
Self-Reliance (SR)	.21	.07		.08	.12	.00	–.15	.11
Dependability (D)	.04	–.05	.08		.00	–.16	–.09	–.17
Sociability (S)	.02	.05	.12	.00		–.11	–.06	–.05
Agreeableness (A)	.04	–.13	.00	–.16	–.11		.16	–.19
Social Perceptiveness (SP)	.29	–.18	–.15	–.09	–.06	.16		–.18
Team Orientation (TO)	–.05	–.04	.11	–.17	–.05	–.19	–.18	
Mean Absolute Difference	.09	.07	.09	.07	.05	.10	.14	.11
Median Absolute Difference	.05	.07	.11	.08	.05	.13	.16	.11

Note. The unit-unique trait weighting scheme was used. The value in each cell represents the Form A correlation minus the Form B correlation. The values in the lower left triangle are repeated in the upper right triangle. Depending upon the sample size for a particular cell, the critical difference in r (at $p = .05$) is between .13–.15.

Select Items and Options of the Concurrent Validation (CV) Version of the PSJT

We sought to develop a concurrent validation version of the PSJT (PSJT-CV)—one test form with 26 items. Each item would have four response options, not seven. Our strategy for developing the PSJT-CV had two important principles. First, we started by identifying psychometrically sound response options and then in turn moving to the item level to identify items for retention. Second, decisions about options and items were based primarily on the psychometric properties for the judgment score key. We did look at correlations between options and personality scales but used these data as tie-breakers for the selection of an option or item. The important point here is that we did not want to jeopardize the ultimate validity of the instrument by trying to maximize too many different keys at the same time and possibly compromising the usefulness of the judgment key. We gave priority to the judgment key.

Option-level statistics. For each response option, we computed the means and standard deviations of the option scores, the option total score correlation, and the option-personality scale score correlation. Average statistics across response options on the two forms appear in Table 10.11. We transferred data for each response option into an ACCESS database containing all options, items, and data to date. One person reviewed the data and selected options for deletion. An option was eliminated if its option-judgment correlation was near zero. The option-trait correlations were considered in borderline cases where the typical psychometrics did not dictate status of the option.

Table 10.11. Option-Level PSJT Statistics for the Full FT Form

Statistic	Form A	Form B
Number of Items	32	32
Number of Response Options	210	211
Average Option-Judgment Correlation	.23	.23
Average Option-Achievement Orientation Scale Correlation	.34	.32
Average Option-Self-Reliance Scale Correlation	.16	.15
Average Option-Dependability Scale Correlation	.31	.36
Average Option-Sociability Scale Correlation	.24	.21
Average Option-Agreeableness Scale Correlation	.22	.16
Average Option-Social Perceptiveness Scale Correlation	.26	.19
Average Option-Team Orientation Scale Correlation	.21	.36

Item-level statistics. For each of the 32 items on each form, we computed the (a) number of retained options for each item, and the (b) average option-total score correlation for the item. Items with fewer than five retained options were eliminated. Then, we selected the final 26 items for the test form by attempting to balance the performance dimensions for the scenarios and maximize the average option-total score correlation for the item. The resulting 26-item form had the following number of items for each dimension: five Adapting to Changing Conditions, six Relating to Peers, four Self-Management, five Self-Directed Learning, and six Teamwork.

Psychometric Properties of PSJT Scores

Having determined which PSJT-FT options, items, and scoring keys should be retained in PSJT-CV, we created two mini-forms—*a* and *b*. Mini-form *a* contained all of the options/items retained from FT Form A, and mini-form *b* contained all of the options/items retained from FT Form B. Our intent is that the combined mini-forms will be one test form in the concurrent validation (i.e., PSJT-CV). We analyzed them separately because recruits were nested within forms in the field test.

Descriptive statistics. We computed means, standard deviations, and reliability estimates for each retained score on each form, using the Spearman-Brown formula to estimate the reliability of scores on PSJT-CV-length versions of mini-forms *a* and *b*. Those data appear in Table 10.12. The table shows that the judgment score for the final 26-item form would likely have a high reliability of about .92. In contrast, the trait reliabilities would likely be much lower in part because there are fewer scored options for each trait. In particular, two of the trait scores in mini-form *b* have negative reliability estimates—Self-Reliance and Sociability. In these two

scales, one option has a negative option-scale correlation. If the results are similar for the options in the concurrent validation, they will likely be excluded when computing the personality scores. Social Perceptiveness had too few options in mini-form *b* to use in computation of the reliability estimate. When the two forms are merged for concurrent validation, the reliability estimates will be more stable and interpretable.

Table 10.12. Means, Standard Deviations, and Reliability Estimates of PSJT scores

Score	*k*	*M*	*SD*	r_{kk}	$r_{26\ item}$
Mini-Form *a*					
Judgment	48	4.48	0.49	.85	.92
Achievement Orientation	13	4.88	0.68	.60	.87
Self-Reliance	5	4.68	0.81	.26	.61
Dependability	5	4.72	0.87	.22	.61
Sociability	4	5.07	1.04	.49	.77
Agreeableness	4	5.11	1.05	.29	.67
Social Perceptiveness	4	5.08	1.11	.42	.64
Team Orientation	5	5.02	0.94	.46	.82
Mini-Form *b*					
Judgment	56	4.57	0.48	.88	.93
Achievement Orientation	13	4.94	0.70	.67	.88
Self-Reliance	6	4.39	0.69	–.04	.00
Dependability	9	5.26	0.89	.69	.87
Sociability	3	4.73	0.93	–.01	.00
Agreeableness	6	4.98	0.92	.58	.82
Social Perceptiveness	1	4.32	1.95	—	—
Team Orientation	8	5.18	0.87	.64	.85

Note. Judgment scores can range from 0 to 6 whereas personality scale scores can range from 1 to 7. Reliabilities were estimated using coefficient alpha. The $r_{26\ item}$ values were computed using the Spearman-Brown prophecy formula; they represent the estimated reliabilities for a 26-item test (with four options per item).

Tables 10.13 and 10.14 report gender and racial/ethnic subgroup differences in PSJT scores, and Table 10.15 reports correlations among the PSJT scores. Female recruits significantly outscored males on the judgment score and several of the personality scale scores. Hispanic and White Non-Hispanic recruits had similar scores. Black recruits tended to score below Whites on most scales. The difference was significant for three scales. Because of the low number of Blacks (particularly for Mini-form *a)*, the difference had to be fairly large to achieve statistical significance; in addition, it is difficult to draw reliable conclusions based on this small sample.

Table 10.13. PSJT Scores by Gender

		Male		Female	
Score on Each Mini-form (*a* or *b*)	d_{FM}	*M*	*SD*	*M*	*SD*
Judgment a	**.31**	4.43	.50	4.58	0.46
Judgment b	**.47**	4.50	.49	4.73	0.42
Achievement Orientation a	.22	4.83	0.71	4.99	0.62
Achievement Orientation b	.07	4.93	0.71	4.98	0.67
Self-Reliance a	.15	4.64	0.83	4.77	0.77
Self-Reliance b	–.20	4.43	0.69	4.29	0.71

Table 10.13. (Continued)

Score on Each Mini-form (*a* or *b*)	d_{FM}	Male *M*	Male *SD*	Female *M*	Female *SD*
Dependability a	.02	4.73	0.85	4.74	0.90
Dependability b	**.28**	5.18	0.94	5.45	0.72
Sociability a	–.08	5.09	1.08	5.00	0.97
Sociability b	–.05	4.74	0.96	4.69	0.85
Agreeableness a	**.35**	5.00	1.08	5.38	0.94
Agreeableness b	.16	4.93	0.96	5.08	0.83
Social Perceptiveness a	**.29**	4.97	1.13	5.30	1.03
Social Perceptiveness b	**.34**	4.10	1.96	4.78	1.90
Team Orientation a	.24	4.95	0.95	5.18	0.91
Team Orientation b	**.32**	5.10	0.88	5.38	0.84

Note. n_{Male}=181, 174; n_{Female}=82, 80. d_{FM}= Effect size for Female-Male mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of the total group. Referent groups (e.g., Males are listed second in the effect size subscript). Statistically significant effect sizes are bolded, p <.05 (two-tailed). A positive effect size indicates that, on average, the non-referent group performs better in the tests.

Table 10.14. PSJT Scores by Race/Ethnic Group

Score on Each Mini-form (*a* or *b*)	d_{BW}	d_{HW}	White *M*	White *SD*	Black *M*	Black *SD*	White Non-Hispanic *M*	White Non-Hispanic *SD*	Hispanic *M*	Hispanic *SD*
Judgment a	–.47	.17	4.48	0.48	4.26	0.64	4.48	0.47	4.56	0.51
Judgment b	**–.39**	.22	4.61	0.47	4.42	0.48	4.59	0.48	4.69	0.41
Achievement Orientation a	–.16	.26	4.87	0.62	4.77	0.99	4.87	0.62	5.03	0.64
Achievement Orientation b	**–.49**	.21	5.01	0.66	4.69	0.75	4.99	0.67	5.14	0.62
Self-Reliance a	–.41	–.16	4.70	0.78	4.38	1.24	4.73	0.74	4.60	0.90
Self-Reliance b	–.31	–.18	4.42	0.70	4.20	0.79	4.44	0.70	4.32	0.64
Dependability a	–.30	.02	4.75	0.82	4.50	1.29	4.76	0.83	4.78	0.81
Dependability b	–.18	.22	5.32	0.87	5.16	0.93	5.29	0.89	5.49	0.70
Sociability a	–.27	.01	5.05	1.03	4.77	1.39	5.08	1.03	5.10	1.10
Sociability b	**–.49**	–.24	4.80	0.85	4.38	1.19	4.84	0.85	4.63	0.85
Agreeableness a	.10	.33	5.12	1.08	5.23	1.03	5.08	1.09	5.44	0.90
Agreeableness b	.04	.19	4.97	0.89	5.00	1.07	4.95	0.92	5.12	0.62
Social Perceptiveness a	.03	.30	5.06	1.12	5.10	1.37	5.05	1.11	5.38	1.15
Social Perceptiveness b	–.23	–.23	4.41	1.88	3.98	2.26	4.45	1.91	4.00	1.88
Team Orientation a	–.06	–.20	5.03	0.92	4.97	1.32	5.06	0.89	4.88	1.04
Team Orientation b	–.29	**.37**	5.23	0.83	4.99	1.01	5.20	0.85	5.52	0.69

Note. n_{White} = 188, 162. n_{Black} = 26, 43. $n_{\text{White Non-Hipanic}}$ = 164, 143. n_{Hispanic} = 31, 30. d_{BW} = Effect size for Black-White mean difference. d_{HW} = Effect size for Hispanic-White Non-Hispanic mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., White) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, p < .05 (two-tailed).

Table 10.15. Correlations Among PSJT Scores for the Final FT Form

Score	J	AO	SR	D	S	A	SP	TO
Judgment (J)		.65	.03	.75	.17	.42	.25	.74
Achievement Orientation (AO)	.68		.29	.61	.27	.58	-.06	.56
Self-Reliance (SR)	.57	.47		.08	.22	.20	-.18	.04
Dependability (D)	.51	.37	.39		.15	.57	.04	.65
Sociability (S)	.52	.43	.32	.34		.32	-.16	.22
Agreeableness (A)	.60	.33	.22	.27	.29		-.21	.44
Social Perceptiveness (SP)	.66	.50	.37	.31	.38	.44		.10
Team Orientation (TO)	.63	.56	.47	.37	.32	.34	.39	

Note. $n_{MiniFormA} = 264$, $n_{MiniFormB} = 256$. Mini-Form A correlations are in the lower left triangle, Mini-Form B correlations are in the upper right triangle.

Discussion

Conclusions

The procedures and results of the PSJT development led to five primary conclusions, as follows:

1. *The 26-item version PSJT can be expected to yield a highly reliable judgment score in the concurrent validation.* Our estimates suggest that the internal consistency reliability estimates for the judgment score will be in upper .80s or lower .90s.

2. *Response distortion and coaching are not likely to affect the PSJT judgment score.* Results of the faking study showed that respondents who were coached to give higher ratings to "actions that get things done, are considerate of other people, maintain/improve morale, and show integrity and honesty" tended to give higher effectiveness ratings to all options, but when their ratings were scored against the key, the respondents' judgment scores were lower with coaching. We also found that coaching strategies that simply ask the respondent to rate all options "4" or avoid using the extreme ratings can be mitigated by stretching the key.

3. *It is reasonable to use civilian scenarios on the PSJT.* We were concerned that military scenarios on the PSJT might require some tacit knowledge of the Army that applicants could not be expected to have and developed civilian versions of test forms to address this issue. The correlations between military and civilian form pairs (r = .70 to .85, and r_c = .83 to .95) were almost as high as that forms' reliability estimates, suggesting that the judgment score was measuring essentially the same thing on the civilian and military forms.

4. *The PSJT judgment score is based on a reliable key developed by credible SMEs.* Coefficient alpha was .93 for each of the two full forms, it was .85 for Mini-form a and .88 for Mini-form b. Eight NCOs were dropped because of too many missing values or very low agreement with the other NCOs.

5. *The PSJT yields reliable scores for four personality traits: Achievement Orientation, Dependability, Agreeableness, and Team Orientation.* When the two forms are merged in the concurrent validation, some of the other personality keys may also yield reliable scores.

Recommendations

Of course, the primary issue of interest is whether scores from the 26-item, civilian-scenario based version of the PSJT are valid predictors of job performance. This question will be addressed in the concurrent validation.

With regard to the PSJT personality scores, the primary issue is whether they have any construct validity as demonstrated by relationships with other personality variables. Chapter 14 provides correlations between scores from various instruments. The PSJT temperament scales did not appear to relate consistently to scores on other temperament scales designed to measure similar constructs. In addition, the two forms of the PSJT temperament scales often correlated differently with other predictor measures, which suggests that the two forms are measuring different constructs. Even so, some of the PSJT traits are not measured directly by any of the other instruments in the Select21 battery. Future research using a well-researched personality instrument as a marker for the traits we are trying to measure on the PSJT would be a better assessment of construct validity. Because all items will be on the same form during the concurrent validation, the construct validity of the personality scores can be better assessed after collecting the concurrent validation data.

CHAPTER 11: PSYCHOMOTOR TESTS

Teresa L. Russell, David Katkowski, Huy Le, and Rodney L. Rosse
HumRRO

Background

There is good reason to expect that tests of psychomotor ability might increment the validity of the ASVAB for predicting certain aspects of future job performance in entry-level Army MOS (McHenry & Rose, 1986). Prior research has shown psychomotor tests to be good predictors of gunnery performance and certain other criteria. Additionally, the Select21 job analysis data suggested that psychomotor constructs were more important for performance in Close Combat jobs and therefore might be discriminating (Sager, Russell, R. C. Campbell, & Ford, 2005).

The problem that the Army has had in trying to implement psychomotor tests has nothing to do with job-relevance, validity, or psychometrics. These tests typically have a specially designed response pedestal and, in the past, have required special computer operating system features. They required extra start-up costs, and they were not easily portable.

The primary research question for Select21 is whether we can develop psychomotor tests that are mechanically and psychometrically reliable and valid using relatively inexpensive, off-the-shelf equipment that can be transported across platforms—specifically, a commercial joystick. We started with the notion that we should use the Army's Project A psychomotor tests to the extent possible (Peterson, 1987). Table 11.1 lists the four Project A psychomotor tests and the constructs they were designed to measure.

Table 11.1. Psychomotor Tests from the Army's Project A

Test	Description	Construct Measured
Target Tracking 1 (One-Hand Tracking)	The respondent uses a joystick, controlled with one hand, to track the movement of a target.	Psychomotor Precision—the ability to make muscular movements necessary to adjust or position a machine control mechanism. Includes Fleishman's (1967) constructs Rate Control and Control Precision.
Target Shoot Test	The respondent tracks a target with a joystick and fires at the target.	
Target Tracking 2 (Two-Hand Tracking)	The respondent uses vertical and horizontal potentiometers controlled with both hands to track the movement of a target.	Multilimb Coordination—the ability to coordinate the movements of a number of limbs simultaneously. It does not refer to tasks in which trunk movement must be integrated with limb movement.
Cannon Shoot Test	The respondent decides when to fire a shell at a moving target and pushes a response button to fire.	Movement Judgment—the ability to judge the relative speed and direction of one or more moving objects.

Incremental Validity

The bulk of the validation analyses performed during Project A used a composite of the scores on the four tests (J.P. Campbell & Knapp, 2001). To learn which tests might be useful in the Select21 project, we reanalyzed Project A data for two Close Combat MOS, 11B (Infantryman) and 19K (M1 Armor Crewman). We used scores on the Project A MOS-specific rating scales (which are similar in nature to the Select21 MOS-specific rating scales described in Chapter 3) as criterion variables since we have developed similar rating scales for Select21. These scales covered performance areas likely to require psychomotor skills, such as "Use of Weapons." We computed the estimated validity of the eight ASVAB subtests currently in use for predicting the average rating on the MOS-specific rating scales and the estimated validities with a psychomotor test added to the ASVAB and/or the experimental spatial test, Assembling Objects, added. The Target Shoot test yielded .05 and .01 incremental validity over the ASVAB for 11B and 19K, respectively. We found this to be heartening news since we will include MOS-specific ratings in our criterion set for Select21. We also found that Target Shoot added to the ASVAB's validity for predicting the M16 Qualification score for 11Bs. These analyses support the inclusion of the Target Shoot test in the Select21 test battery.

Gunnery Performance

In several studies in the late 1980s, ARI found that the two tracking tests from Project A were useful predictors of gunnery performance (Grafton, Czarnolewski, & Smith, 1988; Smith & Graham, 1987; Smith & Walker, 1987). This led Smith and Walker to conclude that "data have consistently shown that by using these tests of psychomotor and spatial ability, it is possible to select gunners who not only are more proficient, but who also take less time to train and qualify" (p. 652).

Classification Efficiency

Using the Enhanced Computer-Assisted Test (ECAT) and ASVAB subtests, Sager, Peterson, Oppler, Rosse, and Walker (1997) attempted to identify "optimal" test batteries that would maximize absolute validity and potential classification efficiency while minimizing three types of subgroup differences (i.e., white-black [WB], white-hispanic [WH], and male-female [MF]). ECAT subjects were Army, Air Force, and Navy trainees, and the criteria were technical ones (e.g., final school grades, hands-on tests). The authors formed a list of the top 20 potential test batteries according to each index (absolute validity classification efficiency, WB, WH, MF). One-Hand Tracking appeared in 19 of the test batteries designed to maximize classification efficiency; Two-Hand Tracking appeared in all 20. They did not include Target Shoot or Cannon Shoot in their analyses.

Practice Effects

Practice effects are always a concern when the testing apparatus is unfamiliar to examinees, and practice effects have been observed on psychomotor tests (McHenry & Rose, 1986). However, the results of practice effects studies for the Project A tests were inconsistent across studies, and the estimated test-retest reliabilities of the tests remained strong despite changes in score magnitude, indicating that the rank ordering of individuals did not change

substantially with practice (cf. Peterson, 1987; Toquam et al., 1986). Based on these results, we decided to examine practice effects in Select21 research.

Conclusions

We determined that two of the psychomotor tests were desirable for the Select21 project psychometrically and practically, Target Shoot and Target Tracking 1. Target Shoot yielded incremental validity over the ASVAB for predicting MOS-specific criteria for selected MOS. Target Tracking 1 showed potential for classification gains in the ECAT Project. Both tests can be administered with one joystick; a full response pedestal is unnecessary.

General Description of the Psychomotor Tests

On each item of the Target Tracking test, a path consisting of vertical and horizontal line segments appears. A target box appears at the beginning of the path. A crosshair is centered in the box. As the item begins, the target starts to move along the path at a constant rate of speed. The examinee's task is to use a joystick to keep the crosshair centered within the target at all times. The examinee's score on this test is the average distance from the center of the crosshair to the center of target across items. This test has 6 practice items and 18 scored items.

At the beginning of an item on the Target Shoot test, a crosshair appears in the center of the screen and a target box appears at some other location on the screen. The target begins to move about the screen in an unpredictable manner, frequently changing direction. The examinee can control movement of the crosshair using a joystick. The examinee's task is to move the crosshair into the center of the target and press a button on the joystick to "fire" at the target. The examinee must do this before the time limit on each trial is reached. The examinee receives three scores on this test. The first is the percentage of "hits" (i.e., the examinee fires at the target when the crosshair is inside the target box). The second is the average time elapsed from the beginning of the trial until the examinee fires at the target. The third score is the average distance from the center of the crosshair to the center of the target at the time the examinee fires at the target. This test has 3 practice and 30 scored items.

Test Development Summary

Developing the initial version of the psychomotor tests involved three steps: (a) select hardware, (b) develop and try-out test construction and delivery software, and (c) pilot test.

Select Hardware

Based on conversations with researchers who had worked on the Air Force's Learning Abilities Measurement Project (LAMP) and Project A, we decided to use a commercial, gaming joystick for Select21. To be useful for our purposes, the joystick needed to meet several criteria:

- Durability— sturdy enough to withstand transportation and use.
- Usability—convenient to plug into the IBM Thinkpads that the Army had purchased, preferably without adding new cards or ports.
- Accuracy—able to track the target closely.

- Ambidextrousness—a joystick that could be used by either left- or right-handed examinees.
- Economy—a common, inexpensive joystick.

With these criteria in mind, we gathered information about joysticks on the Internet from manufacturers and reviewers and decided to use the Logitech Attack 3.

One undesirable feature of commercial joysticks is the centering détente. Unlike the joysticks used in Project A, commercial gaming joysticks have a relatively large centering spring inside the joystick. It creates resistance. We could only speculate how the centering détente and resistance might affect test scores. The spring provides kinesthetic information to the examinee that was not in the original tests. Some test items appear primarily in the center of the screen where the joystick tension would not be as strong, and others appear away from the middle where the joystick tension would be stronger. Importantly, the resistance might affect measurement of the construct we wanted to measure, Precision and Steadiness. Joysticks without the spring require a delicate but steady touch. Joysticks with the spring require muscular force to act against the resistance. We concluded that the springs should be removed.

Develop Test Construction and Delivery Software

The test construction and delivery software developed for this project performs five functions. They include:

1. Administrative tracking—The software contains a module for administration personnel to use in setting up the work station. This includes inputting and tracking computer and joystick identification codes and subject identification codes.
2. Calibration—Apparatus tests need to be calibrated regularly to ensure that the mechanical components of joysticks or other external devices have not become misaligned or worn. The calibration routine is a very simple check on the joystick readings that should be run when the workstation is set up and at regular intervals.
3. Test presentation—In all, the look and the feel of the tests has been designed to emulate the look and feel of the tests used in Project A. The test presentation package includes a joystick familiarization routine as well as the two tests.
4. Test instruction and item editing—The software contains a module for research staff to use to edit and create items.
5. Data output.

Small Sample Try-Out

We created overlength versions of both tests using all of the old items and adding new ones of varying difficulty and conducted a small-sample try-out of the overlength tests at HumRRO. Thirteen HumRRO and ARI staff numbers participated in the effort. Many of the participants were frustrated by the tests and felt that the tests were too difficult. After the try-out, we found and corrected a systematic error in our item files that made the target's speed too fast on all of the items for both tests. Several HumRRO staff took the corrected version of the tests and found the tests to be more reasonable in terms of difficulty.

Pilot Test Psychomotor Tests

In this section, we present a synopsis of the design and results of a pilot study. Data were collected from 124 new recruits. For a detailed description of the pilot test and its results, see Russell and Katkowski (2004).

Research Design

We conducted a within-subjects practice effects study using the pilot test sample. We included two within-subjects factors—practice blocks and difficulty.

Practice blocks. Given constraints on testing time, we were able to administer tests that were twice as long as those used in Project A. We organized the items into two trials, each the same length as its counterpart Project A test. Each trial had two blocks of items to total four blocks—four 9-item blocks of Target Tracking items and four 15-item blocks of Target Shoot items, as shown in Table 11.2.

Table 11.2. Items Organized into Trials and Blocks

	Trial 1		Trial 2	
Test	Block 1	Block 2	Block 3	Block 4
Target Tracking Test	Items 1-9	Items 10-18	Items 1-9	Items 10-18
Target Shoot Test	Items 1-15	Items 16-30	Items 1-15	Items 16-30

To counteract potential boredom and fatigue effects, we alternated Target Tracking and Target Shoot item blocks so that the examinee took nine Target Tracking items, then 15 Target Shoot items, nine more Target Tracking items, and so on. The items in Blocks 1 and 3 had the same items, and Blocks 2 and 4 had the same items. We gave full instructions and practice items for the test the first time it appeared (i.e., six practice items on Target Tracking and three practice items on Target Shoot). Subsequent blocks had minimal instructions and no practice items. After examinees completed the two tests, they answered four questions about their experience playing computer games and using a joystick.

Difficulty. Within the blocks, items were organized according to increasing difficulty. Across the blocks, items were balanced for expected difficulty. For the Target Tracking Test, we used the original 18 Project A items which had been balanced for difficulty. Target Tracking difficulty had nine levels that were a function of two parameters: the speed of the target and the total length of the path of the target. There was one item in each block at each level of difficulty. Target speed and path length increased uniformly across items—the easiest item had the slowest target speed and the shortest path; the hardest item had the fastest target speed and the longest path, making the effects of the two parameters inseparable.

Target Shoot is a more complex test than Target Tracking; consequently a number of factors were included in the difficulty index including (a) the distance from the start of the target to the start of the crosshair (the crosshair always starts in center of the Target Shoot screen); (b) mean segment length, target speed, crosshair speed; and (c) the number of turns. Number of turns, target speed, mean segment length, and crosshair speed are parameters shown to affect test scores in previous Project A work. We included the distance parameter in the difficulty index because, by design, the further away the target began from the crosshair, the larger the minimum latency score.

More time is required for the examinee to navigate the crosshair to the target to shoot when the target begins further away. The difficulty index for Target Shoot ranged from 1 to 6.

Results of the Pilot Test

Target Tracking Test. The pilot test resulted in the following five key findings relevant to the construction of the field test version of the Target Tracking test:

1. Practice had a statistically significant effect on the Target Tracking Test Distance Score.
2. Most of the gains from practice occurred early, during the first block of items. Little to no gain occurred between later blocks of items.
3. The rank ordering of examinees' scores did not change much during the administration of a block of items; internal consistency estimates were .94 or higher *within* the 9-item blocks.
4. The rank ordering of examinees did not change much with practice. The raw correlation between scores on the first and last blocks of 9 items was high (r = .80).
5. With the caveats that the sample size was small (n = 119) and the only other abilities measured were cognitive ones, our data suggested that the abilities that contribute to test performance appear to be stable over the course of practice blocks. Specifically, relationships between Target Tracking Distance Scores and ASVAB scores did not appear to change much with practice.

The Target Tracking Test Distance Scores on individual items are actually the means of many scores, taken every 50 milliseconds during the item. Consequently, the reliability estimate for this test is very high (i.e., r = .98 with 18 items). While there was a practice effect, it occurred early and did not appear to affect examinees' standing relative to each other or the relationship between the Target Tracking Test Distance Score and ASVAB scores.

Target Shoot Test. The Target Shoot Test is more complex than the Target Tracking Test. Many more parameters affect the difficulty of its items, and it yields three potentially useful scores—Hit/Miss, Time-to-Fire, and Distance. Hit/Miss and Distance are accuracy scores, and Time-to-Fire is a speed score.

The Time-to-Fire Score was more reliable than the other two Target Shoot Test scores. The key findings relevant to the construction of the field test version of Target Shoot were as follows:

1. Practice had a statistically significant effect on the Target Shoot test scores.
2. Most of the gains from practice occurred early, during the first block of items. Even so, the effects of practice were not consistent across scores and blocks. For example, there was a substantial gain in Time-to-Fire Scores in the last two blocks of items and relatively little gain for the Distance Score.

3. The rank ordering of examinees' scores did not change much during the administration of a block of items for the Time-to-Fire Score, but internal consistency estimates for the two less reliable scores (Hit/Miss and Distance) tended to increase over the course of the practice blocks.
4. The rank ordering of examinees was somewhat stable regardless of practice. The raw correlation between scores on the first and last blocks of 15 items was .48, .54, and .60 for Hit/Miss, Distance, and Time-to-Fire respectively.
5. With the caveats that the sample size was small (n = 119) and the only other abilities measured were cognitive ones, our data suggested that the abilities that contribute to test performance appear to be stable over the course of practice blocks. Specifically, relationships between Target Shoot and ASVAB scores did not appear to change much with practice.
6. Practice appears to have had a stronger effect on the two, less reliable accuracy scores (i.e., Hit/Miss and Distance). They become more reliable over the course of practice blocks and yield, on average, greater correlations with ASVAB scores on the last block of items.
7. There was a significant gender by practice interaction on two Target Shoot variables. This could indicate that the test was too easy for males (i.e., that there is a ceiling effect for males).

In sum, practice did affect performance on the Target Shoot test, but the effects of practice were inconsistent across the three scores and across practice blocks.

Experience playing video games. During the Project A field test, respondents were asked the following question regarding game-playing experience:

> In the last couple of years, how often have you played video games on arcade machines, home video games, or home computers?
>
> 1. Never
> 2. Less than once a month
> 3. Several times a month
> 4. Once or twice a week
> 5. Almost every day

We asked the same question of our pilot test examinees. The results were highly comparable across studies. In the Project A field test, this question had a mean of 2.99 with an *SD* of 1.03 (n = 256), compared to a mean of 2.96 and *SD* of 1.40 (n = 116) in the Select21 pilot test. As shown in Table 11.3, the correlations between the psychomotor test scores and experience question scores were also comparable in the two studies.

Table 11.3. Correlations Between Psychomotor Test Scores and Video Game-Playing Experience Scores

Score	Project A Field Test n = 250	Select21 Pilot Test n = 116
Target Tracking Distance Score	.22	.21
Target Shoot Hit/Miss Score	--	.21
Target Shoot Time-to-Fire Score	.10	.28
Target Shoot Distance Score	.27	.21

Note. Project A Field Test data are reported in Peterson et al. (1990). Target Tracking had 27 items and Target Shoot had 35. In the Select21 computations, all 60 Target Shoot and 36 Target Tracking items were used.

Recommendations for the field test version. We needed to use administration time in the field test as efficiently as possible. Toward that end, we made several recommendations taking into account the results of the pilot test.

1. *Administer 18 items for Target Tracking.* Given the constraints on time, we decided to give as much time as possible to the Target Shoot Test and make the Target Tracking Test as lean as possible.

2. *Reduce Instruction Time.* In the pilot test, we alternated between tests for blocks of items. Each time the test block changed, refresher instructions for the test were provided. We had switched between blocks to reduce boredom and distribute fatigue and practice evenly across the two tests. We felt this was needed because the Target Tracking items were long and tedious. With a shorter Target Tracking section, it should not be necessary to switch between tests.

3. *Eliminate the Experience Items.* During Project A, the experience items were included in the field test and dropped thereafter. We dropped them from the field test version.

4. *Focus on Target Shoot.* Given that the effects of practice on Target Shoot Test scores were inconsistent and that the Target Shoot Test scores were generally less reliable than the Target Tracking Test Score, we recommended devoting as much of the testing time as possible to the Target Shoot test. We determined that we could continue to examine practice effects in the field test by administering the 30 items twice, if the above changes were made.

Field Test Objectives

Field Test Goals

Our goals for the field test were five-fold:

1. Investigate alternative basic test scores,
2. Examine joystick effects,
3. Examine practice and difficulty effects,
4. Investigate composite scores, and
5. Document psychometric properties of the final scores.

Design

We conducted a within-subjects practice effects investigation using the field test sample, including two within-subjects factors—Trials and Difficulty.

Trials

We administered the 18 Target Tracking once and the 30 Target Shoot test items twice. The 18 Target Tracking items were organized into two 9-item trials, as show in Table 11.4.

Table 11.4. Items Organized into Trials

Test	Trial 1	Trial 2
Target Tracking Test	Items 1-9	Items 10-18
Target Shoot Test	Items 1-30	Items 1-30

Both tests also included three practice items. Examinees took the tests in the following order: (a) three practice items for Target Tracking, (b) 18 Target Tracking items, (c) three practice items for Target Shoot, and (d) two trials of 30 Target Shoot items.

Difficulty

The two Target Tracking trials contained unique items. Items were balanced for difficulty across trials and organized according to increasing difficulty within trials. The two Target Shoot trials contained identical items. Table 11.5 shows the distribution of the 30 Target Shoot items at different difficulty levels.

Table 11.5. Number of Target Shoot Items in Each Trial by Level of Difficulty

Difficulty	Number of Items
1	2
2	4
3	6
4	5
5	7
6	6
Total	30

Field Test Analyses and Results

Data Screening

The psychomotor tests were administered to 663 new recruits in the Select 21 predictor field test. Two of those recruits, were dropped from the sample because their data could not be matched with the Army's records. The sample included 661 cases.

Target Shoot Test

The Target Shoot test requires the new recruit to shoot at the target within a time limit. If the recruit tracks the target, but never shoots, a "no-fire" is recorded, and the recruit has no other

data points for that item. Since low ability examinees who are responding conscientiously to the test may fail to fire at the item, we were concerned about applying a 10% missing data rule without considering the examinee's data. As shown in Table 11.6, 41 of the 661 examinees (6.2% of the sample) would be eliminated from the sample if a 10% missing data rule were applied.

Table 11.6. Number and Percent of "No-Fire" Target Shoot Scores Across 60 Items

Number of No-Fires	Number of People	Percent of Sample
0	264	39.9
1	177	26.8
2	80	12.1
3	46	7.0
4	27	4.1
5	15	2.3
6	11	1.7
7	11	1.7
8	6	.9
9	5	.8
10	4	.6
11	2	.3
12	4	.6
13	2	.3
15	2	.3
16	1	.2
17	1	.2
23	1	.2
27	1	.2
29	1	.2
Total	661	100.0

Note. A 10% missing data rule would eliminate examinees failing to fire on seven or more items.

We screened the sample in four steps. First, we examined the distribution of missing data in Table 11.6 to identify natural breaks in the distribution. Based on those data, we decided to retain examinees with seven or fewer no-fires. Second, we set a cut-point on the amount of data that we would be willing to impute. That is, if we were to retain examinees with missing data, we would have to impute their data. We decided to impute data if the examinee had valid responses for 80% or more of the items (i.e., 48 of 60). Therefore, only examinees with 12 or fewer missing responses would be considered further. At this point, 21 examinees with 8 to 12 missing data points were candidates for inclusion in analyses.

The third step was designed to identify individuals (out of the remaining 21 examinees) who were likely to be responding conscientiously. We classified the Target Shoot items into two groups: easy items (those with difficulty levels lower than or equal to 3) and difficult items (those with difficulty levels higher than 3). There were 24 items (12 for each trial) in the former group and 36 items (18 for each trial) in the latter group. Next, we counted the number of no-fires for each group of items for each the remaining 21 examinees. Finally, for each of the 21 examinees, we used a Chi-square test to determine if the percent of no-fires for items in the

difficult group was significantly higher (one tail test) than that for items in the easy group. Ten of the 21 examinees met this decision rule and were included in the sample. The total sample size after screening was 641.

The fourth step was to review the problem log from the test administrations to identify individuals who appeared to have responded carelessly to the Target Shoot test. This step led to deletion of four more cases. The final sample included 637 cases. This resulted in elimination of approximately 3.6% of the sample (i.e., 24 of 661 examinees).

Target Tracking Test

Only one individual had more than 10% missing data on the Target Tracking test. Indeed that recruit did not have data for any Target Tracking or Target Shoot items and was eliminated from the sample. We deleted 24 examinees who were excluded from the Target Shoot Test sample as described above to simplify tracking of sample sizes in analyses within and between tests.

Joystick and Machine Effects

Joystick and machine effects were a potential source of unwanted error variance in psychomotor test scores. While we did not expect to find either joystick or machine effects since the joysticks and machines were all the same model, we were somewhat concerned that inexpensive commercial joysticks might be somewhat variable in their performance. In the field test, we used 50 different laptops and 29 joysticks. We did not examine laptop effects because the laptops were nested within data collection sites, making systematic analyses difficult. In other words, the laptop effect, if it were to be found, would be completely confounded with test site effect. Since there are many factors that can potentially contribute to the test site effect (e.g., different MOS, composition of male and female), examining such an effect would not be very meaningful.

Joysticks, on the other hand, were used across sites. We examined the joystick effect by conducting a two-way ANOVA with Joystick ID and test site as between subject factors, and Target Tracking Distance Scores as a dependent variable. A significant main effect for joystick would indicate that some joysticks consistently yield higher or lower scores than others. A significant main effect for test site would indicate differential performance by examinees at the two test sites, perhaps due to demographic differences among the new recruits at the sites. We chose to use only the Target Tracking test score as the dependent measure because it is the most reliable of the psychomotor test scores.

Importantly, results of this analysis showed that the joystick main effect was not significant ($p = .23$); therefore, there was no significant difference between the off-the-shelf joysticks. The test site main effect and joystick-by-test site interaction effect were statistically significant ($p < .01$ for both effects).

The significant main effect for test site and the significant joystick by test site interaction are difficult to disentangle and explain. There are huge demographic differences between the sites. The Fort Jackson sample had a large portion of females (i.e., 40%) while Fort Knox had

none. About 72% of the new recruits at Fort Knox were entering Close Combat MOS, and none entering those MOS were at Fort Jackson. Based on prior research with these tests, we would expect the largely male, combat-oriented new recruits at Fort Knox to perform better than those at Fort Jackson (Peterson et al., 1990). Results appear to confirm this expectation; that is, scores from Fort Knox were higher than those from Fort Jackson. If the difference in sites were due to degradation in joystick performance, we might expect the scores from the second site to be lower than those from the first site (i.e., assuming that degradation, wear, and tear result in greater difficulty in tracking the object and therefore lower scores). Since the tests were administered at Fort Jackson first, then at Fort Knox (which yielded higher scores), there is no evidence that this was the case.

Basic Test Scores

We computed the following basic test scores:

- Target Tracking Test Distance Score = Mean of Natural Log (root mean square distance between the crosshair and the target taken at 50 millisecond intervals + 1).

- Target Shoot Test Distance Score = Natural Log (distance between the target and the crosshairs at the "fire" point + 1). If the examinee did not fire at the target, we imputed the Distance Score based on the individual's standing on the test overall. Specifically, we computed z-scores for each item and averaged the z-scores. If Person X did not fire at the target, we computed a Distance Score for that item using the following formula:

$$\text{Imputed Distance}_{\text{Item A}} = \text{MN Distance}_{\text{Item A}} + (SD_{\text{Item A}} * \text{MN Z-Score}_{\text{Person X}})$$

- Target Shoot Hit/Miss Score = If the examinee did not fire at or missed the target, the examinee received a score of "1" for the Hit/Miss Score for that item. Hits were scored "2."

- Target Shoot Time-to-Fire (the amount of time in milliseconds from the onset of the item to shooting at the target). If the examinee did not fire at the target, the examinee received a maximum time score for that item—the full length of time that the item appeared on the screen.

- "No fires"= a proportion of the times that the examinee did not fire.

As in Project A, the Distance Scores were converted to natural logs to correct for skewness in the distribution. This is a fairly standard conversion used on these types of tests. It is important to note that the Target Tracking Distance Score is, for each item, the mean of many Distance Scores. The distance between the target and the crosshair is computed every 50 milliseconds. The score on one item is the mean of these Distance Scores. In contrast, the Target Shoot Distance Score is a point estimate. It is the natural log of the distance between the target and the crosshairs at the "fire" point.

Target Tracking Test

We computed the means, standard deviations, and reliability of the Target Tracking Distance Score and compared our data to the Project A data. As shown in Table 11.7, the estimated reliabilities were quite high and were quite consistent across the two trials of items. The reliabilities are also comparable to those reported during Project A. The means, shown in Table 11.8, were however, different. During Project A, the test was administered on a small computer with a small screen. For this project, we used a larger screen and rescaled the items, making the metric incomparable across the two projects.

Target Shoot Test

As shown in Table 11.7, the four alternative Target Shoot Test scores are quite different in terms of reliability. The latency score, Mean Time-to-Fire, has been the most reliable of the scores in both the Select21 and Project A samples. The remaining three scores are all accuracy measures, with Mean Distance being the most reliable.

Table 11.8 provides the means and standard deviations of the alternative Target Shoot scores, and Table 11.9 provides the correlations among them. The means and standard deviations of the Project A scores are also provided, although they are not directly comparable to the Select21 scores due to differences in the size of the computer screen and the items. Select21 means and score correlations are discussed in further detail in the next section on practice and difficulty effects.

Based on the reliability estimates and correlations, we decided to retain Mean Time-to-Fire and Mean Distance for further analysis. We retained Time-to-Fire because it is the most reliable score. We retained the Mean Distance Score because it was the most reliable of the three accuracy scores, was highly correlated with the Hit-Miss score (r = .89), and yielded a low correlation with Time-to-Fire (r = .09).

The No-Fire score has a relatively low reliability (.64) due to the low base-rates of No-Fires. Nevertheless, we retained this score for possible use in special-purpose analyses. It is possible that the decision whether to fire is a function of personality as well as psychomotor ability.

Table 11.7. Reliability Estimates for Alternative Psychomotor Test Scores

	Odd-Even Split-Half Reliability Estimates[a]								Test-Retest Estimates		
	Select21 Field Test[b]			Select21 Pilot Test[c]			Project A[d]		Project A CV[d]	Select21 Pilot Test[c]	Select21 Field Test[b]
	Trial 1	Trial 2	All	Trial 1	Trial 2	All	Incumbents (CV)	New Recruits (LV)			
Target Tracking											
# Items	9	9	18	18	18	36	18	18	18	18	9
Distance Score	.97	.96	.98	.97	.97	.97	.98	.98	.74	.87	.94[e]
Target Shoot											
# Items	30	30	60	30	30	60	30	30	30	30	30
Hit/Miss Score	.65	.70	.78	.75	.65	.81	.52	.66	.44	.69	.58
Mean Time-to-Fire	.91	.91	.95	.88	.83	.92	.85	.84	.58	.81	.77
Mean Distance Score	.75	.81	.86	.76	.73	.85	.74	.73	.37	67	.64
Number of No-Fires	.51	.44	.64	.70	.54	.72	--	--	--	--	.46

[a]Trial reliability estimates were corrected to total number of items in the trial using the Spearman-Brown correction.
[b]*n* for the Select21 Field Test was 637. The tests were administered with no delay interval.
[c]*n* for the Select21 Pilot Test was 119. The tests were administered with no delay interval.
[d]Project A data are provided for two samples. The concurrent validation (CV) sample consisted of incumbents and the longitudinal validation sample (LV) was composed of new recruits. *n*s for the Project A samples on which the split-half reliability estimates were computed were 9099-9274 (CV) and 6436 (LV). The *n* for the CV test-retest data was 473-479. The test-retest interval for the CV data was one month. LV data are from Peterson et al. (1990), and CV data are from Toquam et al. (1986).
[e]Adjusted to 18 items using the Spearman-Brown equation.

Table 11.8. Means and Standard Deviations of Psychomotor Test Scores by Trial

		Select21 Field Test[a]			Select21 Pilot Test[b]			Project A[c]	
Test Score		Trial 1	Trial 2	All	Trial 1	Trial 2	All	Incumbents	New Recruits
Target Tracking									
# Items		9	9	18	18	18	36	18	18
Distance Score	*M*	3.95	3.83	3.89	3.91	3.73	3.82	2.98	2.89
	SD	.51	.52	.50	.49	.55	.50	.49	.46
Target Shoot									
# Items		30	30	60	30	30	60	30	30
Hit/Miss Score[d]	*M*	1.60	1.65	1.62	1.56	1.65	1.60	.53[d]	--
	SD	.14	.15	.13	.15	.16	.14	.13	--
Time-to-fire	*M*	412.15	408.69	410.42	407.06	388.95	398.01	235.39	230.98
	SD	122.38	113.77	111.15	127.76	107.92	112.08	47.78	50.18
Distance Score	*M*	2.67	2.61	2.64	2.73	2.61	2.67	2.17	2.20
	SD	.23	.28	.23	.24	.26	.23	.24	.23
Number of No-Fires	*M*	.026	.020	.023	.035	.017	.026	--	--
	SD	.041	.034	.032	.052	.034	.037	--	--

[a] n = 637 for the Select21 Field Test sample.
[b] n = 119 for the Select21 Pilot Test sample.
[c] *ns* for the Project A samples on which the means and standard deviations were computed were 9099-9274 (incumbents in the concurrent validation) and 6436 (new recruits in the longitudinal validation). Data for the Project A samples are from Peterson et al. (1990) and Toquam et al. (1986).
[d] Proportion of Hits to Misses. Not computed for the new recruit sample.
Note. Time-to-Fire Scores are in hundredths of seconds. Distance Scores are in natural log units.

Table 11.9. Correlations Among Alternative Target Shoot Scores

	Score			
Score	Hit-Miss	Time-to-Fire	Distance	Number No-Fires
Hit-Miss	*.78*			
Time-to-Fire	.22	*.95*		
Distance	.89	.09	*.86*	
Number No-Fires	.34	.75	.25	*.64*

Note. Split half reliability estimates appear on the diagonal. Time-to-Fire, Distance, and No-Fires scores were reversed so that those who performed better in the tests had higher scores. n = 637.

Practice and Item Difficulty Effects

We performed two-way ANOVAs to investigate the effects of practice and difficulty on Target Tracking and Target Shoot (Distance and Time-to-Fire) test scores. We expected the main effects for both factors to be significant based on prior research (McHenry & Rose, 1988; Peterson, 1987; Peterson et al., 1990).

Main Effect of Difficulty

The main effect for difficulty was significant in all three ANOVAs. As shown in Figure 11.1, the Target Tracking Distance Score increases as the difficulty of the item increases. The graph also illustrates the main effect of practice. Distance Scores are larger (i.e., worse) for the first trial than they are for the second one. Similarly, Figures 11.2 and 11.3 illustrate the main effect of difficulty for the two Target Shoot scores. Generally, as difficulty increases, the Time-to-Fire Score increases (i.e., it takes longer to shoot), and the Distance Score increases (i.e., the crosshairs are further from the target when the shot is fired).

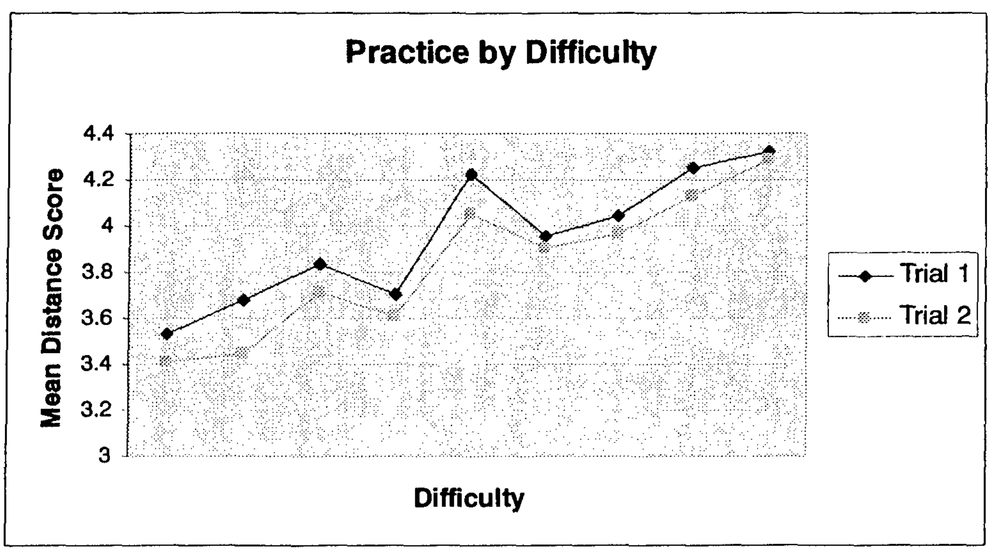

Figure 11.1. Target Tracking Test Distance Score by difficulty and trial.

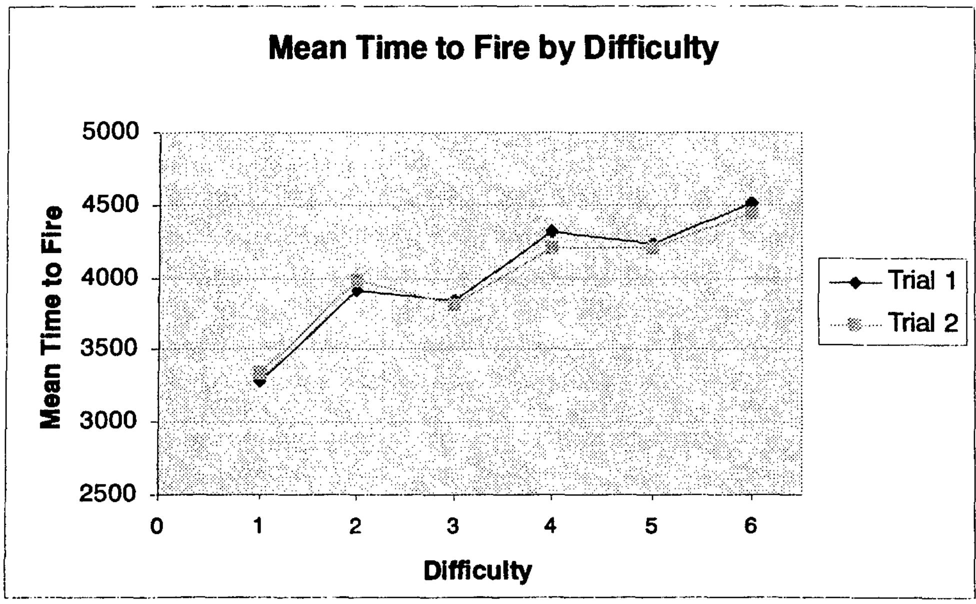

Figure 11.2. Target Shoot Test Time-to-Fire Score by difficulty and trial.

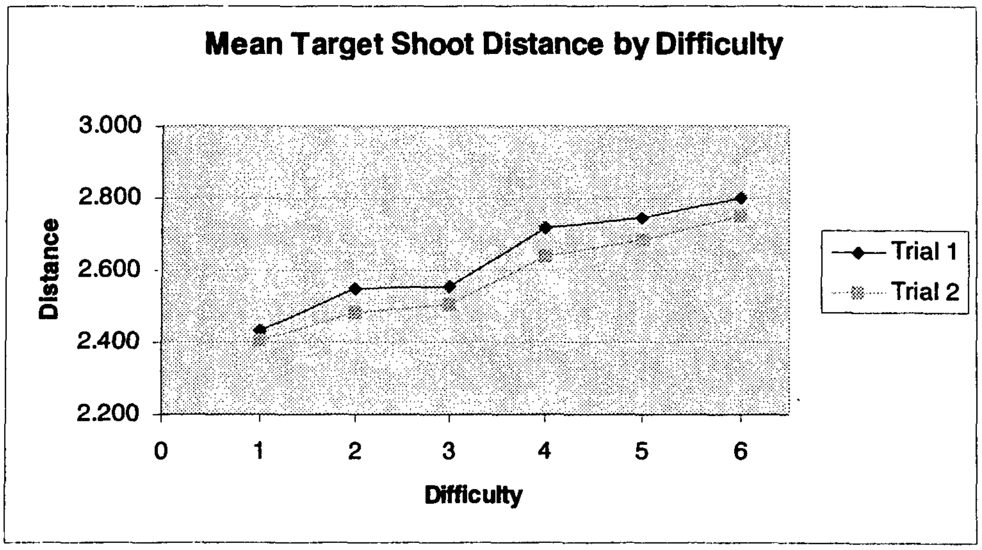

Figure 11.3. Target Shoot Test Distance Score by difficulty and trial.

Main Effect of Practice (Trial)

Practice effects studies routinely show gains of one-third of an *SD* or more for cognitive ability tests (Russell, Reynolds, & J. P. Campbell, 1994) as well as psychomotor tests (McHenry & Rose, 1988). For example, General Aptitude Test Battery (GATB) researchers conducted 11 practice effects studies, with a total of 2783 subjects (Department of Labor, 1970). The average gain in group mean scores from the first testing to the second testing, in standard deviation units, was .81 for finger dexterity, .91 for manual dexterity, and .45 for motor coordination, compared to effect sizes ranging from .31 to .55 for the cognitive GATB aptitudes.

Accuracy or distance scores. In the Select 21 field test and pilot test, the main effect of practice was significant for the two accuracy scores, Target Tracking Distance and Target Shoot Distance. As shown in Table 11.10, examinees gained more than 1/3 *SD* on the distance scores in the pilot test and about 1/4 *SD* on the same scores in the field test.[45] These results are consistent with those from Project A. In the Project A field test practice effects research (Peterson, 1987), pre- and post-practice testing occurred 2 weeks apart. Practice included retesting on new items and occurred about one week after the initial test. A control group also took the pre- and post-tests. The practice group gained about 1/3 of an *SD* on the two distance scores; the control group also improved, but less so. Later, during concurrent validation, the psychomotor tests were a part of a test-retest study. The test-retest interval was one month, with no practice. Toquam et al. (1986) reported a gain of .27 *SD* on the Target Tracking Distance Score and no gain on the Target Shoot Distance Score.

Latency or Time-to-Fire Score. The main effect of practice was not significant for the Target Shoot Time-to-Fire score.[46] This result is consistent with the results of the Project A in that Toquam (1986) reported that the examinees performed slightly worse at retest than they had during the initial testing. (The Time-to-Fire score was not computed during the Project A practice effects research.)

Rank ordering of individuals. Even though the practice effects were significant for the distance scores, the test-retest correlations (between trial 1 and trial 2) for these scores were relatively high (Table 11.7) suggesting that practice did not have a strong effect on the rank ordering of individuals' scores.

Correlations with ASVAB Subtests

If learning occurs on a test, it is possible that the abilities that contribute to test performance change over the course of many trials. For example, a series of studies (Ackerman, 1986, 1987, 1988, 1989) investigated the abilities that predict performance on controlled processing tasks (i.e., tasks that require conscious effort to apply rules in new situations). They found that the sets of abilities that predict performance change when the tasks have been practiced to the point of reaching automaticity, effortless performance.

We computed the correlations between the psychomotor test scores and ASVAB scores by trial. For the Target Tracking Distance Scores, the rank orders of the correlations with ASVAB subtest scores are highly stable across blocks, with the highest correlation always being with Mechanical Comprehension scores and the lowest always being with Paragraph Comprehension or Math Knowledge scores. There is no evidence of change in abilities related to the Target Tracking scores with practice. Scores on the Target Shoot test (Distance and Time-to-Fire) like those on the Target Tracking test yield a fairly consistent pattern of correlations across trials, again with the highest correlation usually being with Mechanical Comprehension scores and the lowest with Paragraph Comprehension or Math Knowledge scores.

[45] The differences in gain scores between the pilot and field tests are probably due to the differences in size and composition of the two samples. The pilot test sample was small and predominately female, while the larger field test sample was mostly male.
[46] The main effect of practice was significant for Time-to-Fire in the Select21 pilot test.

Table 11.10. Gain Scores from Practice

	Project A Practice Effects				Project A Test-Retest		Select21 Pilot Test		Select21 Field Test	
	n	Two-Week Gain Score with Practice	*n*	Two-Week Gain Score without Practice	*n*	One-Month Gain Score without Practice	*n*	Back-to-Back Administration Gain Score	*n*	Back-to-Back Administration Gain Score
Target Tracking Distance Score	74	.33	113	.07	487	.27	119	.34	637	.23
Target Shoot Time-to-Fire Score	--	--	--	--	468	-.04	119	.14	637	.03
Target Shoot Distance Score	74	.26	113	.21	468	.00	119	.50	637	.21

Note. Gain scores are expressed in standard deviation units. Positive gain scores indicate better performance at retest. The test-retest interval for the Project A field test was 2 weeks (Peterson, 1987). Only Target Shoot Distance Scores were analyzed in the Project A field test practice effects study. The test-retest interval for Project A was one month. There was no test-retest interval in the Select21 pilot test and field test. Tests were administered back-to-back.

Further Tests for the Learning Effects

To examine factors that potentially influenced changes in the psychomotor test scores across trials, we conducted further analyses based on regression models with change scores as the dependent variable (i.e., the difference between the individual's trial 1 and trial 2 scores) and gender and cognitive abilities (measured by the ASVAB) as the independent variables (cf. Judd, Kenny, & McClelland, 2001). Two groups of regression models were tested. Model 1 looked at the effect of gender and general cognitive ability (as measured by the AFQT) on the change scores. Model 2 was meant to examine the effects of more specific abilities, so gender and the ASVAB subtests were included as the independent variables.

Effect of cognitive ability on distance or accuracy change scores. As can be seen in Table 11.11, neither of the two distance change scores (i.e., Target Tracking and Target Shoot) were affected by cognitive abilities. In Model 1, the beta weight for AFQT was not significant, and none of the cognitive ability weights in Model 2 were significant.

Table 11.11. Effects of Gender and Abilities on the Changes of Psychomotor Test Scores Across Trials

Independent Variables	Dependent Variables (Change in Psychomotor Test Scores)		
	Target Tracking Distance Score	Target Shoot Distance Score	Target Shoot Time-to-Fire Score
	Std. Beta Weight	Std. Beta Weight	Std. Beta Weight
Model 1: Gender and AFQT			
Gender	-.01	.09**	.03
AFQT score	-.02	.03	-.08 *
Model 2: Gender and the ASVAB Sub-tests			
Gender	-06	.10 *	.00
General Science	.06	.04	-.19**
Arithmetic Reasoning	-.05	-.06	-.02
Word Knowledge	.09	-.04	.12
Paragraph Comprehension	.00	.03	-.03
Math Knowledge	.02	.11	.02
Electronics Information	-.08	-.04	-.04
Auto-Shop Information	-.04	.07	-.03
Mechanical Comprehension	-.05	.05	.04
Assembling Objects	-.02	-.02	.02

Note. *$p < .05$; **$p < .01$; Female = 1; Male = 0; Positive change scores indicate improvement in test performance.

Effect of cognitive ability on latency or time-to-fire change scores. As for the Time-to-Fire score, AFQT and the General Science subtest score had significantly negative effects on the change score. That means people who have higher cognitive abilities tended to improve *less* across trials than those with lower cognitive abilities. We have no good explanation for this

finding except to say that cognitive ability might be related to changes with practice on the latency score.

Effect of gender. Explanations of gender differences sometimes involve the differential experience hypothesis (Sherman, 1967)—the idea that women have had less experience with certain tasks than men and therefore have not reached the asymptote of their ability. If so, with practice, women, being lower on the learning curve, should improve more than men. Alternatively, a significant beta weight for gender could indicate that a psychomotor test is too easy for males. That is, males' performance would be hindered by a ceiling effect. While females' scores would improve, males' scores would remain roughly the same.

Gender was significant for one of the three scores, the Target Shoot Distance Change Score. Females tended to improve their scores more through practice as compared to males. During the pilot test, we had conducted gender by practice ANOVAs and found that the gender by practice interaction was significant for this test score and not the others. At the time, we were concerned that the Target Shoot test could be too easy for males. Since this test is most likely to be used for classification into combat jobs, it is important that it be appropriate for males. Gender differences could be a moot point. Before the field test, we replaced three of the easier items with more difficult ones in case the test was too easy.

To examine the possibility of a ceiling effect in the field test data for the Target Shoot Distance Score, we reasoned that the following four conditions would suggest a ceiling effect:

1. The distribution of males' scores would be skewed compared to the female score distribution.
2. The *SD* of males' scores would be smaller than the *SD* of females' scores.
3. Females would improve their scores more between trials as compared to males. That is, the gender by trial interaction would be significant.
4. There would be not much difference in males' scores across difficulty levels; whereas, females would perform better on easier items (i.e., difficulty levels of 1 or 2) and worse on more difficult items (difficulty levels of 5 or 6). This would be demonstrated by a significant gender by difficulty interaction.

After assessing these conditions, we concluded that there was no clear evidence of a ceiling effect. With regard to Condition 1, we examined the graphs of distributions for both Target Shoot Distance and Time-to-Fire scores and the male score distribution did not appear to be more skewed than the female score distribution. Also, the *SD* of males' Target Shoot Distance Scores (.21) was about the same as that for females (.23) (cf. Table 11.20). We conducted a three-way ANOVA (difficulty, trial, and gender, with difficulty and trial as within-group factors and gender as a between group factor) to address Conditions 3 and 4 using the Target Shoot Distance Score as the dependent variable. Condition 3 was met. Females improved more than males. But, Condition 4 was not met; the gender by difficulty interaction was not significant.

Development of Composite Scores

We considered three different ways to combine the three basic scores (i.e., Target Tracking Distance, Target Shoot Distance, and Target Shoot Time-to-Fire). One method was to keep the Target Tracking Distance Score by itself and form a composite of the two Target Shoot scores. Even though the two Target Shoot scores are not correlated with each other ($r = .09$, Table 11.9), it could be argued that the bottom line of performance on the test is shooting accurately and quickly at the target. The downside is that the psychometric evidence suggests that the two scores are getting at different constructs. They are reliable scores, uncorrelated with each other, and affected differentially by practice (i.e., Time-to-Fire scores did not improve with practice).

A second method was to combine all three scores into a general psychomotor composite. This method could be a desirable way of preserving degrees of freedom in validation regression analyses. Even so, it suffers from the same psychometric problems described above.

We chose a third method—to create composite by adding the standardized distance Scores for the two tests. We labeled it Psychomotor Precision. In this case, the latency (Time-to-Fire) score remains as a separate score. The empirical rationale was that the two distance scores were correlated with each ($r = .51$), and both improved with practice while the Time-to-Fire score did not. There was also support for this decision on the theoretical side. The two distance scores were originally intended to tap Fleishman's (1967) two accuracy constructs, Rate Control and Control Precision (Peterson, 1987). The Time-to-Fire score was added by the Project A team, and the team was not quite sure how this score fit in the psychomotor domain (Peterson, 1987). In exploratory factor analyses, it yielded split loadings on two factors, General Psychomotor and Perceptual Speed (which was defined by decision time scores on perceptual speed tests), with the loading on the General Psychomotor factor being slightly higher than the other loading. Ultimately, the Time-to-Fire score was merged, along with all of the psychomotor test scores, into one psychomotor composite.

Table 11.12 provides the descriptive statistics for the psychomotor basic scores and the composite score. It also provides correlations among them. The estimated reliabilities of the basic test scores appear in Table 11.7. The reliability of the Psychomotor Precision composite, estimated using Feldt and Brennans's (1989) formula for the reliability of a composite score, was .95.

Table 11.12. Descriptive Statistics for and Correlations among Psychomotor Test Scores

	Total Group		Correlations			
Psychomotor Score	*M*	*SD*	TTD	TSD	TSTF	PP
Target Tracking Distance (TTD)	3.89	.50	--			
Target Shoot Distance (TSD)	2.64	.23	.51	--		
Target Shoot Time-to-Fire (TSTF)	410.42	111.48	.53	.09	--	
Psychomotor Precision Composite (PP)	.00	1.75	.87	.87	.37	--

Note. $n = 641$.

Subgroup Differences

Tables 11.13 and 11.14 provide scores for gender and racial subgroups. The effect sizes are comparable in magnitude to those obtained during Project A (Peterson et al., 1990).

Table 11.13. Psychomotor Test Scores by Gender

		Male		Female	
Psychomotor Score	d_{FM}	*M*	*SD*	*M*	*SD*
Target Tracking Distance	**.88**	3.77	.47	4.21	.42
Target Shoot Distance	**.70**	2.59	.21	2.75	.23
Target Shoot Time-to-Fire	**.67**	387.37	106.95	462.00	103.78
Psychomotor Precision Composite	**.95**	-.48	1.55	1.19	1.63

Note. n_{Male}=442, n_{Female}=178, d_{FM}= Effect size for Female-Male mean difference. Effect sizes calculated as (mean of non-referent group-mean of referent group)/*SD* of the total group. Referent groups (e.g., Males) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, p <.05 (two-tailed). A positive effect size indicates that in average the referent group performs better in the tests.

Table 11.14. Psychomotor Test Scores by Race/Ethnic Group

			White		Black		White Non-Hispanic		Hispanic	
Psychomotor Test Scores	d_{BW}	d_{HW}	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Target Tracking Distance	**.88**	**.24**	3.80	.45	4.24	.55	3.80	.44	3.92	.52
Target Shoot Distance	**.26**	-.09	2.63	.22	2.69	.23	2.64	.22	2.62	.20
Target Shoot Time-to-Fire	**.70**	**.26**	393.87	108.01	471.72	110.08	391.42	108.40	420.74	110.23
Psychomotor Precision Composite	**.67**	.10	-.21	1.56	.97	1.89	-.19	1.53	-.02	1.82

Note. n_{White} = 412. n_{Black} = 99. $n_{White\ Non\text{-}Hipanic}$ = 356. $n_{Hispanic}$ = 80. d_{BW} = Effect size for Black-White mean difference. d_{HW} = Effect size for Hispanic-White Non-Hispanic mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., White) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, p < .05 (two-tailed).

We also examined differences between scores of recruits in a wide assortment of MOS (Army-Wide) compared to those assigned to Close Combat (CC) MOS. As shown in Table 11.15, new recruits for CC MOS performed better than those recruited for AW MOS. Since the CC MOS have no females, we thought the effect might simply be due to gender difference. We examined the differences in psychomotor test scores between AW and CC MOS in the sub-sample consisting of only male participants. Table 11.15 shows the results. As can be seen, most of the differences, though reduced in magnitude, remained significant. This result suggests that there may be indirect range restriction in the sample of participants in the Close Combat MOS. In other words, new recruits might have been selected into CC MOS based on variables that are correlated with psychomotor abilities. The pattern of differences in standard deviations between the Army-wide and Close-Combat MOS samples reported in Table 11.18 (i.e., standard deviations of psychomotor test scores of the Close Combat MOS sample are smaller than those of the Army-wide sample) is indeed consistent with this speculation. However, we could not directly test this hypothesis.

Table 11.15. Descriptive Statistics for Psychomotor Test Scores by MOS

	Total Group			Army-Wide		Close-Combat	
Psychomotor Score	*M*	*SD*	d_{AC}	*M*	*SD*	*M*	*SD*
Target Track Distance	3.89	.50	**-.60**	3.95	.50	3.65	.38
Target Shoot Distance	2.64	.23	**-.26**	2.65	.22	2.59	.24
Target Shoot Time-to-Fire	410.42	111.48	**-.45**	419.02	108.85	368.93	104.21
Psychomotor Precision Composite	.00	1.75	**-.53**	.18	1.79	-.75	1.36

Note. n_{AW}=488, n_{CC}=131, d_{FM}= Effect size for Close-Combat-Armywide MOS mean difference. Effect sizes calculated as (mean of non-referent group-mean of referent group)/*SD* of the total group. Referent groups (e.g., Army-wide MOS) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, p <.05 (two-tailed). A positive effect size indicates that in average the referent group performs better in the tests.

Table 11.16. Descriptive Statistics for Psychomotor Test Scores by MOS – Males Only

	Total Group			Army-Wide		Close Combat	
Psychomotor Score	*M*	*SD*	d_{AC}	*M*	*SD*	*M*	*SD*
Target Track Distance	3.77	.47	**-40**	3.82	.50	3.63	.36
Target Shoot Distance	2.59	.21	-.10	2.60	.20	2.58	.24
Target Shoot Time-to-Fire	387.37	106.95	**-.27**	395.32	105.77	366.21	103.34
Psychomotor Precision Composite	-.48	1.55	**-.28**	-.35	1.63	-.79	1.32

Note. n_{AW}=305, n_{CC}=128, d_{FM}= Effect size for Close (Combat-Army-wide MOS mean difference. Effect sizes calculated as (mean of non-referent group-mean of referent group)/*SD* of the total group. Referent groups (e.g., Army-wide MOS) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, p <.05 (two-tailed). A positive effect size indicates that in average the referent group performs better in the tests.

Conclusions and Recommendations

The primary purpose of the field test was to obtain psychometric data and practice effect information that would inform the development of the concurrent validation versions of the Select21 psychomotor tests.

Conclusions

We drew four main conclusions from the analyses. They are as follows:

1. The psychomotor tests yielded reliable scores that are likely to be useful for the concurrent validation, particularly for use in classification analyses.
2. The joysticks appear to be relatively comparable to each other in terms of test scores they produce. The main effect of joystick was not significant.
3. While there is a practice effect on these tests, we do not find it to be alarming or of great concern. The psychomotor test scores did improve with practice, perhaps by one quarter to one-third of an *SD*, and improvements of that magnitude are often observed for cognitive tests (Russell et al., 1994). Additionally, the rank ordering of examinees' scores did not change much during the administration of a block of items, as indicated by reasonably high internal consistency estimates. The data suggested that the abilities that contribute to test performance appear to

be stable over the course of practice blocks. Specifically, relationships between test scores and ASVAB scores did not appear to change much with practice.

4. The data and prior research suggested that it would be reasonable to combine the two distance scores to form a composite (Psychomotor Precision) and retain the latency (Time-to-Fire) score remains as a separate score. The empirical rationale was that the two distance Scores were correlated with each (r = .51), and both improved with practice while the Time-to-Fire score did not. The two distance Scores were originally intended to tap Fleishman's (1967) two accuracy constructs, Rate Control and Control Precision (Peterson, 1987).

Recommendations for the Concurrent Validation

We recommend using the psychomotor tests in the concurrent validation research. We expect that psychomotor scores could be particularly useful for classification into close combat MOS.

We examined test completion rates and estimated time requirements and internal consistency reliability estimates for different versions of the tests with different lengths. If we have 25 minutes for administration of this these two tests during concurrent validation, we recommend including 52 Target Shoot items and nine Target Tracking items. The reliability estimates for each of the three basic scores should be .85 or greater at this test length. If the tests must be shortened for a 20 minute administration, we recommend including 40 Target Shoot items and eight Target Tracking items. In this case, all of the scores should retain reliabilities of .80.

CHAPTER 12: RECORD OF PRE-ENLISTMENT TRAINING AND EXPERIENCE (REPETE)

Teresa L. Russell, Huy Le, and Deirdre J. Knapp
HumRRO

Background

Historically, the Army has taken the burden of training all required entry-level job skills for its enlisted personnel. It stands to reason that recognizing prior related training and/or experience could benefit the Army by reducing training requirements (or at least helping to ensure success in training) and benefit applicants by enhancing their enlistment options (in terms of job choices and/or enlistment bonuses). With much the same reasoning, the Air Force investigated military-relevant skills of enlistees in 1971 and estimated that approximately 40% of personnel entering the Services had job skills likely to transfer to military jobs (Hoehn, Wilson, & Richards, 1972). They concluded that taking advantage of prior training and experience could have substantial payoff for the Services. Even so, the Services have not conducted research on self-report experiential measures in recent years.

The Record of Pre-Enlistment Training and Experience (REPETE) is a self-report measure designed to determine the type of training and experience that entry-level Soldiers currently bring with them to the Army. It might be used in non-traditional ways (e.g., to allow recruits to "test out" of particular training courses, to provide enlistment bonuses for particular experiences); therefore, it is not necessarily a predictor in the traditional sense. Because the REPETE asks for entry-level experiences, it is not suitable for concurrent validation. Soldiers with 18 to 36 months of experience in the Army would have to respond retrospectively. These factors suggested that the development of the REPETE would best be characterized as a demonstration project—one designed to illustrate what kind of instrument might be developed and how it might be used—rather than a predictor development effort.

For the purposes of the demonstration effort, we focused on basic computer skills. Computer skills are somewhat important for all jobs and particularly important for some MOS (Sager, Russell, R. C. Campbell, & Ford 2005), but they are not addressed directly on the ASVAB as are some other skills (e.g., electronics, auto/shop). The volatility of the computer industry has made it difficult to include computer skills in traditional measurement tools. These factors suggested that attempting to develop a measure of computer experience would be a worthwhile demonstration effort.

Development of REPETE Content Domains

We developed a preliminary taxonomy of computer skill domains by identifying and reviewing research on computer skills, particularly test content definitions used by high school testing programs (Bradlow, Hoch, & Hutchinson, 2002) and the Army (Dyer & Martin, 1999). We also content analyzed and sorted computer-related community/technical school courses (e.g., Saint Paul College) and certification testing programs (e.g., Brainbench). We cross-referenced the sorts/taxonomies from different sources and identified common themes. This process resulted in 10 content domains.

We used data collected from roughly 600 new recruits at Fort Benning and Fort Jackson in the fall of 2003 to refine the basic computer skills content domains and to identify other, potentially useful certification content areas. We asked the new recruits to list, in an open-ended fashion, certifications obtained beyond high school.

A fairly large number of respondents wrote in names of courses that they had taken in high school and wrote the name of their high school as the certifying body. Several wrote in the names of scholarships or awards they had received. Others wrote in their job titles and names of employers as certifying organizations. If the write-in response appeared to be a high school course or a job rather than a certification, we did not include it in our analysis.

We sorted the responses from Forts Benning and Jackson, and Table 12.1 summarizes the result. There were three main findings. First, the computer categories seemed to hold up pretty well. Based on these results, we added a category for networking/computer service support, dropped file management, added graphics to presentation software, and added desktop publishing to word processing software. Table 12.2 provides the final 10 computer categories.

The second main finding was the large number of write-ins in several other areas, particularly medical and protective service areas as shown in the table. After reviewing the data, we decided to supplement the computer content areas with six categories of certifications for the field test version of the REPETE. The six categories and their definitions appear in Table 12.3.

General Description of the Field Test Version of the REPETE

The field test version of the REPETE has three multi-part items. In the first item, respondents are asked to list courses they have taken related to computer skills and to indicate which of the 10 categories were addressed in each course. The second item is structured similarly, but asks about certifications. The second item has a second part that asks respondents to list other certifications they have in the six additional areas listed in Table 12.3. The last item asks respondents to rate themselves on each of the 10 computer categories using the following 5-point level-of-mastery rating scale:

1= Little or no skill in this area
2= I am familiar with the basics in this area
3= I have a solid working knowledge and skill in the basics of this area
4= I have knowledge and experience in some advanced concepts in this area
5= I am an expert in this area, highly experienced with the most advanced applications

Field Testing of the REPETE

The primary objectives of the field testing of the REPETE were two-fold:

- to identify and evaluate alternative scoring schemes and determine whether the three item types appear to provide complementary data (i.e., information from the three item types are neither negatively nor very highly positively correlated), and
- to content-analyze the write-in responses to the REPETE.

Table 12.1. Summary of Write-In Responses from Fort Benning and Fort Jackson

Type of Certification	Certification	Number of Responses
Business License	Insurance	2
	Mortgage Broker	1
Home Services	Babysitter (Red Cross)	2
Computer	Word Processing	9
	Spreadsheet Software	4
	Database Software	1
	Internet Usage	1
	Presentations/Graphics	4
	Basic Hardware and Operating Systems	5
	Computer Programming Principles	2
	Basic Web Programming	2
	Computer Networking Service and Support	3
	File Management	0
	Object Oriented Programming	0
Driver's License/ Certification	Boat	5
	Truck	2
	motorcycle/snowmobile	2
	cab	1
	Other (forklift, cherry picker)	8
	Flight Ground School	1
Fitness	Scuba diving, swimming, and snorkeling	12
	Martial Arts	1
	Skydiving	1
Food Service	Food Handler's Permit (and just food service)	4
	Chef	1
Mechanical	Automobile Repair and Maintenance (e.g., ASE Certified Technician, AUTOCAD 2000)	16
	ASE Certified Parts SALES	1
Medical	Certified Nurses Assistant	10
	Certified Nurse Practitioner	1
	Emergency Medical Technician	10
	CPR	115
	Lifesaving (usually Red Cross)	36
	First Aid	32
	Phlebotomy	4
	Dental something	1
	Physical Therapy	2
	X-ray	1
	Home health aide	1
	Defibrillator	1
	Others	6
Protective Service	Firefighter, Fire safety related	17
	Police- Related/ Security guard cert.	18
	Firearm Certification/ Permit to Carry	17
	Hazardous Materials	3
	Hunter's license	4
Skilled Trades	Electrical	4
	Carpentry	3
	Welding	4
	Plumbing	1
	Painting	1
	HVAC	1
Other	Certified Public Speaking	1
	Tutoring	1
	Teaching	1
	Guitar Instructor	1
	Minister's License	1
	Public Relations (Red Cross)	1

Table 12.2. Basic Computer Skills Categories

1.	Word Processing/Desktop Publishing Software—Create, manipulate, format, and print documents.
2.	Spreadsheet Software—Record, format, sort, analyze, and graph information.
3.	Database Software—Create, query, organize, analyze, graph, and report databases.
4.	Presentation/ Graphics Software—Create presentation-quality slides or graphics.
5.	Internet Usage and Information Search—Send, receive, and save email; search the internet.
6.	Basic Hardware and Operating Systems—Manage own pc files and folders using the operating system and hardware.
7.	Networking and Computer Service Support—Install, initialize, configure, and manage network software
8.	Computer Programming Principles—Develop algorithms, select programming languages, design program, use assembly language, and develop documentation.
9.	Basic Web Programming—Develop web pages using HTML and Javascript.
10.	Object-Oriented Programming Concepts—Create Object-Oriented Modeling using UML notation; use C++, Java, or Visual Basic.

Table 12.3. Six Supplemental REPETE Skill Areas

1.	Health Services and Medical Skills—Skills in emergency care and medical care and technology.
2.	Protective Service Skills—Skills in handling firearms, and activities such as police and fire that involve protecting the public.
3.	Mechanical Skills—Skills in repair and maintenance of mechanical equipment
4.	Electrical Skills—Skills in the installation, repair, and maintenance of electrical systems.
5.	Driving and Piloting Skills—Skills in driving or piloting vehicles or heavy equipment.
6.	Athletic Skills—Skills in athletic pursuits such as swimming, martial arts, scuba diving, and skydiving.

Identify Scoring Schemes

Descriptive Statistics

One concern regarding the REPETE is the potentially low base rate for items asking Soldiers to list courses taken and certifications obtained. We computed frequency distributions for the numbers of courses taken and certifications obtained to examine those base rates.

More than half of the respondents (53%) had taken one computer course (Table 12.4), and another 22% had taken two. But, the base rate on computer certifications obtained was very low. Table 12.4 also shows that only 48 respondents had obtained one computer certification (8%), and very few had obtained more than one. Table 12.5 provides tallies of the number of courses and certifications respondents' indicated were relevant to the 10 computer skills categories.

Table 12.4. Number of Computer Courses and Computer Certifications

Number of Computer Courses	Respondents		Number of Computer Certifications	Respondents	
	n	%		*n*	%
0	94	15.3	0	539	90.3
1	327	53.3	1	48	8.0
2	135	22.0	2	8	1.3
3	27	4.4	3	0	0.0
4	14	2.3	4	1	0.2
5	11	1.8	5	1	0.2
6	4	0.7			
7	0	0.0			
8	1	0.2			
Total	613	100.0		597	100.0

There were also very low base rates for the numbers of other certifications as shown in Table 12.6. Fewer than 5% of the respondents had obtained certifications in the protective service, mechanical, and electrical categories.

Table 12.7 provides the means and standard deviations of the scores of all of the variables, including the responses on the 5-point level of mastery rating scale. As shown in Table 12.7, we also computed three summary variables: (a) total number of computer certifications, (b) total number of computer courses, and (c) mean mastery rating on computer courses.

Tables 12.5 and 12.7 show very low response rates of certifications addressing Computer Programming and Object Oriented Programming skills. These skills were also rated lowest in terms of self-assessed levels of mastery, and were the least likely to be addressed in courses taken by the participants Hence, we decided to drop Computer Programming and Object Oriented Programming skills from further analyses.

Composite Scores

If individuals who take certain courses are also likely to obtain certifications and rate themselves higher in those areas, it would be reasonable to combine these three variables into more reliable composite scores. Table 12.8 provides the correlations among the computer-related variables. As shown, all of the correlations were positive, although their magnitudes were rather modest probably because of the skewed distributions of the Certifications and Courses variables. In general, correlations between Courses and Mastery variables (mean r = .41) were higher than those involving Certifications (mean r = .23 and .25 for correlations between Certifications and Courses, and Certifications and Mastery, respectively). This finding suggested that Courses and Mastery variables were likely to reflect similar constructs, and combining these two variables would result in more internally consistent composites (as compared to forming composites with the Certifications variables).

Table 12.5. Number of Computer Courses (Co) and Certifications (Ce) for Each Computer Skill Category

Number	1. Word-processing/ Desktop Publishing Software		2. Spreadsheet Software		3. Database Software		4. Presentation/ Graphics Software		5. Internet Usage and Information Search		6. Basic Hardware and Operating Systems		7. Networking and Computer Service Support		8. Computer Programming Principles		9. Basic Web Programming		10. Object-Oriented Programming	
	Co	Ce	Co	Ce	Co	Ce	Co	Ce	Co	Ce	Co	Ce	Co	Ce	Co	Ce	Co	Ce	Co	Ce
0	184	564	291	574	362	575	328	579	305	571	378	571	500	578	523	588	461	583	528	588
1	345	32	260	23	206	21	224	18	252	23	187	22	88	15	70	8	127	12	65	8
2	67	1	50		38	1	51		49	3	36	2	21	2	16	1	22	2	16	1
3	8		8		3		7		4		7	1	2	0	2		1		1	
4	4		2		1		1		0		2	1	1	2	2		2		3	
5	4		2		3		2		3		3		1							
6	1																			
7																				
8																				
9																				
10																				
Total	613	597	613	597	613	597	613	597	613	597	613	597	613	597	613	597	613	597	613	597

Note. Co=Courses; Ce=Certifications. Each number in a cell of the table is a count of the number of times the examinee indicated a course or certification was relevant to the computer skill category.

Table 12.6. Certifications Obtained in Each of Six Supplemental Areas

Number of Certifications	Health & Medical		Protective Service		Mechanical		Electrical		Driving & Piloting		Athletics	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
0	526	85.8	585	95.4	590	96.2	603	98.4	553	90.2	521	85.0
1	48	7.8	21	3.4	18	2.9	9	1.5	53	8.6	55	9.0
2	27	4.4	6	1.0	2	0.3	1	0.2	5	0.8	19	3.1
3	7	1.1	0	0.0	2	0.3			2	0.3	9	1.5
4	0	0.0	0	0.0	0	0.0					7	1.1
5	5	0.8	1	0.2	1	0.2					2	0.3
Total	613	100.0	613	100.0	613	100.0	613	100.0	613	100.0	613	100.0

Table 12.7. Means and SD of All REPETE Variables

Variable	*M*	*SD*
Number of Courses: Word Processing	.89	.81
Number of Courses: Spreadsheet Software	.66	.76
Number of Courses: Database	.51	.72
Number of Courses: Presentation	.59	.75
Number of Courses: Internet	.62	.73
Number of Courses: Basic Hardware	.49	.76
Number of Courses: Networking	.24	.57
Number of Courses: Computer Programming	.19	.51
Number of Courses: Web Programming	.30	.57
Number of Courses: Object Oriented Programming	.18	.52
Total Number of Computer Courses	1.34	1.08
Number of Certifications: Word Processing	.06	.24
Number of Certifications: Spreadsheet Software	.04	.19
Number of Certifications: Database	.04	.20
Number of Certifications: Presentation	.03	.17
Number of Certifications: Internet	.05	.24
Number of Certifications: Basic Hardware	.06	.30
Number of Certifications: Networking	.05	.30
Number of Certifications: Computer Programming	.02	.14
Number of Certifications: Web Programming	.03	.18
Number of Certifications: Object Oriented Programming	.02	.14
Total Number of Computer Certifications	.12	.43
# Health & Medical Certifications	.24	.71
# Protective Service Certifications	.06	.33
# Mechanical Certifications	.05	.33
# Electrical Certifications	.02	.15
# Driving & Piloting Certifications	.11	.37
# Athletic Certifications	.26	.74
Mastery: Word Processing	2.92	1.18
Mastery: Spreadsheet Software	2.39	1.11
Mastery: Database	2.08	1.05
Mastery: Presentation	2.50	1.30

Table 12.7. (Continued)

Variable	*M*	*SD*
Mastery: Internet	3.67	1.21
Mastery: Basic Hardware	2.65	1.38
Mastery: Networking	1.88	1.14
Mastery: Computer Programming	1.49	0.88
Mastery: Web Programming	1.82	1.09
Mastery: Object Oriented Programming	1.39	0.81
Mean Mastery: Computer Skills	2.28	0.83

Note. The frequency variables are counts of the number of courses or certifications that examinees listed to open-ended questions. The mastery variables are respondents' self-ratings on a 5-point level of mastery scale.

Table 12.8. Correlations Among Computer-Related Variables

	Correlation Coefficient			
Computer Skill Category	Certifications/ Courses	Certifications/ Mastery	Courses/ Mastery	Mean
1. Word-processing & Desktop Publishing	.13	.17	.24	.18
2. Spreadsheet Software	.16	.16	.37	.23
3. Database Software	.24	.23	.41	.29
4. Presentation/Graphics Software	.17	.20	.50	.29
5. Internet Usage & Information Search	.21	.17	.30	.23
6. Basic Hardware & Operating Systems	.33	.21	.32	.29
7. Networking & Computer Service Support	.43	.32	.49	.41
8. Computer Programming Principles	.14	.27	.44	.28
9. Basic Web Programming	.26	.36	.49	.37
10. Object-Oriented Programming Concepts	.23	.39	.53	.38
Overall	.18	.31	.31	.27
Mean Correlation	.23	.25	.41	.30

Note. The Overall row provides the following three correlations and their mean: (a) Total number of certifications with total number of courses, (b) total number of certifications with mean mastery across 10 categories, and (c) total number of courses with mean mastery across 10 categories.

We further explored the possibility of combining the variables by carrying out a separate exploratory factor analysis (EFA) on each of these three types of variables: Courses, Certifications, and Mastery. Results of these analyses indicated that there was one factor underlying the eight Mastery variables.[47] Similar results were found for the Courses variables. For Certifications, it appeared that there were two correlated factors underlying the eight computer skill variables. The first four computer skills (i.e., Word Processing, Spreadsheet, Database, and Presentation) had the highest loadings on the first factor. The remaining four skills (i.e., Internet, Basic Hardware, Networking, and Web-Programming) loaded on the second factor.[48] We called these factors Basic Computer Skills (the first factor) and Advanced Computer Skills (the second factor).

[47] We used the eigenvalues > 0 criterion and examined the resulting scree plot to determine the number of factors to be retained.

[48] It can be argued that these factors are just artifacts due to the difference in response frequencies. That is, the first four skills tended to be addressed more often in the Certifications obtained by the participants, whereas the last four skills were less likely to be addressed. Even if this is true, we believe that this difference has substantive meaning in the sense that they reflect the underlying nature of the variables (basic computer skills vs. advanced computer skills). Thus, it makes sense to treat these factors separately.

We created a total of six composite scores for use in additional analyses. Three composite scores were summary variables reported in Table 12.7:

1. Total number of computer certifications.
2. Total number of computer courses.
3. Mean mastery rating on computer courses.

Based on the correlational and EFA results described above, we created the following three composite scores:

4. General Computer Skills – the sum of standardized scores of all the 16 Courses and Mastery variables.
5. Basic Computer Certifications – the sum of four Certifications variables: Word Processing, Spreadsheet, Database, and Presentation.
6. Advanced Computer Certifications – the sum of the four remaining Certifications variables: Internet, Basic Hardware, Networking, and Web-Programming.

Descriptive Statistics for Composite Scores

Table 12.9 provides descriptive statistics for and correlations among the six REPETE composite scores. We estimated the reliability of the composite scores using coefficient alpha.[49] Those estimates appear in the diagonal of Table 12.9. Tables 12.10 and 12.11 provide descriptive statistics for the composite scores by gender and race respectively.

Table 12.9. Descriptive Statistics for and Correlations among REPETE Composite Scores

	Total Group		Correlations					
Composite Scores	*M*	*SD*	1	2	3	4	5	6
1. Total Number of Computer Courses	1.34	1.08	--					
2. Total Number of Computer Certifications	0.12	0.43	.18	--				
3. Mean Mastery: Computer Skills	2.28	0.83	.31	.31	*.90*			
4. General Computer Skills	0.12	10.26	.52	.30	.86	*.90*		
5. Basic Computer Certifications	0.16	0.70	.12	.64	.30	.29	*.88*	
6. Advanced Computer Certifications	0.18	0.90	.14	.80	.31	.30	.58	*.89*

Note. Reliability estimates appear on the diagonal of the correlation matrix.

Content Analysis of Write-In Responses

Future versions of the REPETE would have a checklist format. To begin development of that format, we content-analyzed the write-in responses to REPETE questions about certifications and courses. The goal of the content analysis was to produce descriptors that would provide official certification names and computer courses that are likely to be taken by the applicant population. The first step was to identify identical or almost identical certification or

[49] Reliability of the General Computer Skills composite was estimated based on the variances/covariances of eight combined items, which were formed by summing the standardized scores on the respective Courses and Mastery variables.

course names. Next, we researched websites of certifying organizations (e.g., Microsoft, National Institute of Automotive Service Excellence) to find official certification names and course catalogs from three community colleges (Cuyahoga Community College in Cleveland, OH, St. Paul College in St. Paul, MN, and Merced Community College in Merced, CA) to identify relatively common names and descriptions for the courses taken.

Table 12.10. REPETE Composite Scores by Gender

		Male		Female	
Composite Score	d_{FM}	*M*	*SD*	*M*	*SD*
Total Number of Computer Courses [a]	**-0.21**	1.41	1.19	1.18	0.78
Total Number of Computer Certifications [b]	0.00	0.12	0.46	0.12	0.38
Mean Mastery: Computer Skills [c]	-0.05	2.29	0.87	2.25	0.71
General Computer Skills [d]	0.01	0.05	10.82	0.13	8.72
Basic Computer Certifications [b]	-.001	0.17	0.72	0.16	0.64
Advanced Computer Certifications [b]	**-0.17**	0.22	1.03	0.07	0.39

Note. [a] $n_{Male} = 425$, $n_{Female} = 184$. [b] $n_{Male} = 414$, $n_{Female} = 179$. [c] $n_{Male} = 384$, $n_{Female} = 166$. [d] $n_{Male} = 388$, $n_{Female} = 167$. d_{FM}= Effect size for Female-Male mean difference. Effect sizes calculated as (mean of non-referent group-mean of referent group)/*SD* of the total group. Referent groups (e.g., Males) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

Table 12.11. REPETE Composite Scores by Race/Ethnic Group

			White		Black		White Non-Hispanic		Hispanic	
Composite Scores	d_{BW}	d_{HW}	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Total Number of Computer Courses [a]	**.25**	-.06	1.30	1.03	1.57	1.29	1.29	1.04	1.23	0.89
Total Number of Computer Certifications [b]	.07	-.12	0.11	0.45	0.14	0.37	0.12	0.48	0.07	0.26
Mean Mastery: Computer Skills [c]	.18	-.22	2.26	0.81	2.41	0.89	2.27	0.82	2.09	0.76
General Computer Skills [d]	.19	.00	-0.32	10.10	1.61	10.24	-0.50	10.07	-0.50	10.35
Basic Computer Certifications [b]	.20	-.07	0.13	0.60	0.27	0.91	0.14	0.64	0.09	0.50
Advanced Computer Certifications [b]	.02	-.09	0.16	0.94	0.18	0.70	0.18	1.00	0.10	0.54

Note. [a] $n_{White} = 398$, $n_{Black} = 99$, $n_{WhiteNon\text{-}Hispanic} = 352$, $n_{Hispanic} = 70$. [b] $n_{White} = 387$, $n_{Black} = 96$, $n_{WhiteNonHispanic} = 341$, $n_{Hispanic} = 70$. [c] $n_{White} = 363$, $n_{Black} = 87$, $n_{WhiteNonHispanic} = 323$, $n_{Hispanic} = 62$. [d] $n_{White} = 366$, $n_{Black} = 88$, $n_{WhiteNonHispanic} = 325$, $n_{Hispanic} = 63$. d_{BW}= Effect size for Black-White mean difference. d_{HW}= Effect size for Hispanic-White non-Hispanic mean difference. Effect sizes calculated as (mean of non-referent group-mean of referent group)/*SD* of the total group. Referent groups (e.g., Whites) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

Computer Courses

As indicated in Table 12.3, more than half of the respondents had written in at least one computer course name. The content analysis of their responses appears in Table 12.12. As shown, the most frequently taken courses were Computer Applications I/Computer Fundamentals, Keyboarding, and Word Processing. A few individuals had taken higher level programming and network administration courses.

Table 12.12. Content Analysis of Write-In Responses for Computer Courses

n	Course
227	Computer Applications I / Computer Fundamentals
86	Keyboarding, Typing, and Data Entry I
42	Word Processing
35	Management/Business Computer Information Systems
25	Computer Science I
24	Computer Applications II
21	Microcomputer Hardware/Software Maintenance, Repair, and Support
21	Web Page Design with HTML/JavaScript
18	Computer Programming Basics
18	Network Administrator/Associate Course 1
15	Computer Graphics: Basics
13	Computer-Aided Design/Drafting
10	Desktop Publishing I
10	Spreadsheets: Microsoft Excel
7	Computer Graphics: Digital Design and Imaging
7	Computer Programming: C/C++
6	Computer Science II
6	Microsoft PowerPoint
5	Visual Communication
4	Database Use and Design
4	Internet Fundamentals
4	Microsoft Windows
4	Visual Communication: Media Design
3	Computer Programming: Java
2	Computer Programming: Numerical Analysis
2	Computer Programming: Visual Basic
2	Visual Communication: Digital Video
1	Accounting Computer Applications
1	Computer Graphics: 2D Animation and Video
1	Computer Graphics: 3D Modeling and Rendering
1	Computer Programming: Assembly Language
1	Computer Programming: COBOL
1	Computer Programming: Fortran
1	Desktop Publishing II
1	Keyboarding, Typing, and Data Entry II

Computer Certifications

As shown in Table 12.4, few respondents wrote-in computer certifications. Most of those write-ins were for basic topics such as Computer Fundamentals and MS Excel Fundamentals (Table 12.13). However, some of the certifications (e.g., A+, Cisco Certified Network Professional, Microsoft Certified Professional) are advanced; indeed, many jobs in the computer world, such as network administrator, require one or more of these certifications for job entry.

Table 12.13. Content Analysis of Write-In Responses for Computer Certifications

n	Response Content	Example Certifying Organization
8	Computer Fundamentals	Brainbench
7	A+	Computing Technology Industry Association (CompTIA)
5	MS Excel Fundamentals	Brainbench
5	MS Word Fundamentals	Brainbench
3	Auto CAD	Society of Automotive Engineers
3	Cisco Certified Network Professional	Cisco
2	Microsoft Certified Professional	Microsoft
2	Microsoft Certified Systems Engineer	Microsoft
2	Network + Certified Service Technician	Computing Technology Industry Association (CompTIA)
2	Networking Concepts	Brainbench
2	MS Office Specialist	Microsoft
2	Web Design Concepts	Brainbench
1	Electronic Switching System (ESS)	Enterasys
1	IBM Authorized Repair Center Technician	IBM
1	Information Technology Terminology	Brainbench
1	Microsoft Certified Applications Developer	Microsoft
1	Microsoft Certified Systems Administrator	Microsoft
1	MS Access Fundamentals	Brainbench
1	MS Windows Fundamentals	Brainbench
1	Web Hosting	Brinkster

Certification in Six Supplemental Skill Areas

The write-ins for the other certifications listed a wide range of individual achievements, and are provided in Appendix G. In general, there were few write-ins for the six supplemental skills.

Many of the respondents who wrote in items for "Athletic Certifications" wrote in items we had not intended to elicit, such as "Most Valuable Player." The bulk of the write-in responses referred to an achievement in a particular sport (n = 80), such as winning a state Rodeo competition, not a certification per se. A few respondents indicated certification or licensure in scuba diving or skydiving.

The most frequent response for "Driving and Piloting Certifications" (n = 36) was having a driver's license. A couple of respondents had licenses for commercial vehicle, aircraft, or industrial equipment operation. The most frequent response for protective service-related certifications was Firearms Certification or Licensure (n = 6). Similarly, there were very few write-ins for the mechanic-related certifications, with Certified Autobody Repair Technician (n = 4) being the most prevalent, and there were no coherent electrical certification write-ins.

Compared to the other supplemental skill areas, responses for "Health Care Certifications," seemed to have more potential for a future REPETE. The most frequent response was CPR and First Aid, as shown in Table 12.14. The other responses, while infrequent, represent medical skills that might be of interest to the Army.

Table 12.14. Content Analysis of Write-In Responses for Health Care Certifications

n	Content
78	CPR and First Aid
12	Certified Nursing Assistant (CNA)
5	Emergency Medical Technician (EMT)
3	Certified Phlebotomist
1	Cardiac-Related: Defibrulator Certification
1	Cardiac-Related: Life Support
1	Cardiac-Related: Vascular Technician

Recommendations for Future Research

The Select21 field testing of the REPETE showed strong response rates for computer courses, demonstrated that reasonably reliable composite scores can be created, and importantly showed that some Army recruits have coursework, certifications or licenses that are likely to be MOS-relevant and represent advanced work in an area. Conceptually, these findings suggest that the REPETE could be useful for (a) classifying applicants into MOS, (b) identifying individuals who might be able to "test out" of introductory MOS-specific training, or perhaps (c) awarding enlistment bonuses.

The REPETE will not be included in the concurrent validation research for Select21 because the participants (job incumbents) would have to respond retrospectively, in terms of what they had done prior to entering the Army, and we are concerned about the accuracy of the retrospective response. If the Army is interested in pursuing further development of the REPETE, the first step would be to determine the likely uses and purposes. In turn, the purpose would drive the development of content for the new REPETE. For example, if the purpose is to attract individuals with specific computer or medical skills by providing an enlistment bonus, the next version of the REPETE would list specific courses and certifications of interest for those skills. Once content is defined, the rational decisions about how to weight and score the REPETE would need to be made.

If the REPETE is to be used for computer skill assessment, decisions will also need to be made about how to handle the volatility of the computer industry. There are literally hundreds of certifications offered in the computer world (cf., Brainbench.com), but many of them are very narrow (e.g., MS Word 3.0, MS Word 2000). Because they are narrow, they also change with each new generation of software. That would make the REPETE difficult to maintain. One possibility would be to focus only on the more advanced industry certifications. Another would be to eliminate certifications altogether because much of the same information comes from the computer courses. For example, community colleges offered courses in Microcomputer Hardware/Software Maintenance, Repair, and Support that were designed to prepare students for A+ certification.

CHAPTER 13: PERSON-ENVIRONMENT FIT MEASURES

Chad H. Van Iddekinge, Dan J. Putka, and Christopher E. Sager
HumRRO

BACKGROUND

Personnel selection measures are typically designed to assess the knowledge, skills, and attributes (KSAs) critical to performance in the job of interest. Although important, job performance is not the only criterion of concern to most organizations. For example, organizations like the U.S. Army are interested in reducing attrition through personnel selection and classification. Traditional KSA-based measures, however, are seldom designed to predict both performance and alternative criteria such as attrition (Hom & Griffeth, 1995).

In recent years, personnel researchers have turned to measures of person-environment (P-E) fit to predict outcomes other than job performance. Considerable research has demonstrated that scores on such measures are often related to various work-related attitudes and intentions (e.g., job satisfaction, organizational commitment, turnover intentions), as well as to behaviors such as absenteeism and turnover (e.g., Cable & DeRue, 2002; Saks & Ashforth, 1997; Verquer, Beehr, & Wagner, 2003). In this chapter, we describe several predictor measures being developed to assess fit with regard to the current and future Army work environment. We begin by providing some background on developing predictors of alternative criteria. We then discuss the P-E fit predictors we developed and the results of analyses evaluating their psychometric characteristics and potential effectiveness within the Army context.

Developing Predictors of Alternative Criteria

The P-E fit measures described in this chapter were designed to predict first-term attrition and the attitudinal criteria discussed in Chapter 7. The two main theoretical frameworks that influenced the development of these measures are discussed below.

The theory of work adjustment (TWA; Dawis, England, & Lofquist, 1964) provided the primary theoretical foundation for the Select21 fit measures. TWA suggests that job satisfaction is a function of the correspondence between workers' preferences for various *occupational reinforcers* and the degree to which the job or organization provides those reinforcers.[50] An occupational reinforcer is a characteristic of the work environment associated with an individual's work-related needs (e.g., having a chance to be creative, being paid well, having good peer relationships). Correspondence between workers' needs and the needs the job or organization "supplies" is referred to as *needs-supplies fit* (Edwards, 1991; Kristof, 1996). Measures of needs-supplies fit are the most common type of P-E fit measure, and as discussed, have been shown to predict turnover and its attitudinal precursors (e.g., job satisfaction). Thus, one class of measures we describe in this chapter is designed to assess fit between applicants' needs and the Army work environment.

---

[50] Holland's (1985) congruence theory makes a similar prediction regarding the correspondence between individuals' vocational interests and the interests a work environment supports.

A second theoretical framework influenced us to develop another class of fit measures, namely those designed to assess fit between applicants' expectations and the realities of the Army work environment. Research suggests that some applicants have unrealistically high and/or inaccurate expectations regarding the work environment (Wanous, 1992). It has been suggested that providing applicants a realistic job preview (RJP) during the recruitment or selection process can bring applicants' expectations more in line with reality and, in turn, reduce later negative effects of unmet expectations (Hom, Griffeth, Palich, & Bracker, 1999). Indeed, there is empirical evidence that the use of RJPs can increase job satisfaction and reduce turnover (e.g., Wanous, 1992). Nonetheless, RJPs are not used as a selection device. In fact, they often leave the selection decision to applicants (i.e., self-selection), which can be a disadvantageous to organizations, particularly in a lean recruiting environment.

One way the Army might capitalize on the effectiveness of RJPs, yet keep the selection decision in the hands of the Army, is to present RJP information in a pre-service "knowledge of the Army" test. For example, prospective recruits could be asked to rate the extent to which the "needs" assessed in a needs-supplies fit measure are characteristic of the Army. In addition to selection, the Army could use such information in a variety of other ways. For instance, recruits with inaccurate expectations could be identified for subsequent pre- or post-enlistment counseling interventions (e.g., during the Delayed Entry Program [DEP]). In light of these possibilities, we developed measures to assess fit between applicants' expectations and the realities of the Army work environment, or what we refer to as *expectations-reality fit.*

Another reason for exploring applicants' expectations in Select21 is that we believe expectations may moderate relations between needs-supplies fit and various criteria. Specifically, we hypothesize that "misfit" between recruits and the Army regarding a particular interest or value (e.g., need for autonomy) depends on (a) how important autonomy is to the recruit, (b) how much autonomy the recruit expects the Army to provide, and (c) how much autonomy the Army actually offers. For example, consider two recruits, one who values autonomy and expects the Army to supply it, and a second who values autonomy, but does not expect the Army to supply it. If the Army does not supply autonomy, it is likely that the second recruit will be more satisfied in the Army than the first. Although both recruits value autonomy (indicating a lack of needs-supplies fit), the fact that the first recruit expects autonomy and does not receive it may result in greater dissatisfaction for the first recruit.

Thus, needs-supplies and expectations-reality fit are the two types of fit that theory and empirical research indicate are most likely to predict attrition and the attitudinal criterion measures we developed. Nevertheless, we are also exploring the utility of assessing *abilities-demands fit*, or the correspondence between KSAs applicants possess and the KSAs required by a job or organization (Edwards, 1991; Kristof, 1996). Within the fit literature, abilities-demands fit has most often been viewed as a precursor to occupational stress (Edwards, 1996), and Soldier stress, in turn, has been consistently linked to a variety of attrition-related criteria (e.g., Strickland, 2004). Despite the potential effectiveness of measures of abilities-demands fit, we chose not to focus on assessing this type of fit because we felt that such measures would be unlikely to increment the predictive validity of the Select21 KSA-based predictors.[51]

[51] A potential exception is the prediction of perceived stress assessed by the Army Life Survey (ALS).

Nevertheless, in the final section of this chapter we explore abilities-demands fit with regard to the temperament-related requirements of Army work.

Content Domains of the P-E Fit Measures

As discussed, predictor measures are generally designed to assess the critical KSAs identified by a job analysis. However, when developing measures of fit, the constructs critical to assess may vary by applicant (e.g., it depends on what values or interests an applicant finds most desirable) instead of being a fixed set of KSAs. As such, when developing fit measures for Select21, we identified broad taxonomies of work-related needs to cover the range of needs the enlisted applicant population might desire. The results of prior research suggest that needs underlying vocational interests and work values are most relevant to predicting turnover and its attitudinal precursors (Dawis & Lofquist, 1984; Holland, 1985). Given this, interests and values serve as the primary content for the Select21 P-E fit measures. As mentioned, we also explore the use of temperament variables for assessing abilities-demands fit.

Table 13.1 displays the Select21 P-E fit measures by content domain and type of fit. In the following sections, we describe how each measure was developed, the results of relevant data analyses, and any remaining issues or concerns we have. We begin by discussing measures designed to assess fit with regard to vocational interests, and then discuss the work values and temperament content measures. For each set of instruments, we describe the environment-side supplies measure(s), the person-side needs measure(s), and the person-side expectations measure.

Table 13.1. Select21 P-E Fit Measures

	Type of Measure				
	Person-Side			Environment-Side	
Content	Needs Measures	Expectations Measures	Abilities Measures	Supplies Measures	Demands Measures
Vocational Interests	Work Preferences Survey (WPS)	Pre-Service Expectations Survey (PSES)		Army Environment Survey (AES)	
	Interest Finder Questionnaire (IFQ)			Job Characteristics Survey (JCS)	
Work Values	Work Values Inventory (WVI)	Army Beliefs Survey (ABS)		Army Description Inventory (ADI)	
Temperament		Army Work Knowledge Survey (AWKS)	Work Suitability Inventory (WSI)[a]		Work Styles Supply Survey (WSSS)

Note. Needs and supplies measures are used to assess needs-supplies fit, expectations and supplies measures are used to assess expectations-reality fit, abilities and demands measures are used to assess abilities-demands fit, and expectations and demands measures are used to assess expectations-demands fit. We also developed alternate versions of the AES and ADI supplies measures that subject matter experts (SMEs) used to indicate the vocational interests and work values they expect the future Army work environment to support.
[a]Development of this measure is discussed in Chapter 9.

VOCATIONAL INTERESTS MEASURES

The first set of measures is designed to assess fit with regard to vocational interests. We focused on the interest constructs from Holland's (1985) congruence theory. Similar to the TWA, this theory suggests that job satisfaction is a function of the congruence between individuals' work interests and the interests supported by their job or organization. According to Holland, vocational interests are expressions of personality that can be used to categorize individuals and work environments into six types: realistic, investigative, artistic, social, enterprising, and conventional (RIASEC). Holland's model has been widely validated and is the prevailing taxonomy in vocational psychology (Barrick, Mount, & Gupta, 2003). We begin by discussing the "supplies" RIASEC measures, which we will use to generate the Army interests profiles against which we will compare interests profiles of Soldiers.[52]

Environment-Side Supplies Measures

Army Environment Survey

Description of Measure

The Army Environment Survey (AES) is a 24-item instrument that assesses the extent to which the current Army (in general) supports each RIASEC dimension. The purpose of the AES is to create an Army-wide interests profile with which applicant profiles (i.e., based on a needs measure) can be compared to assess P-E fit. The AES was developed after a thorough review of articles and source materials on Holland's model and the general vocational interests literature. NCOs (i.e., Drill Sergeants and Advanced Individual Training [AIT] instructors) served as SMEs and were asked to read a description of each RIASEC dimension and then rate the extent to which it characterizes the Army environment by answering four Likert-type items (e.g., "First-term Soldiers with Realistic interests would be satisfied with the Army."). SMEs made their ratings on a 5-point Likert-type scale with anchors that ranged from *strongly disagree* (1) to *strongly agree* (5). After rating each dimension, SMEs were asked to rank the dimensions in terms of how well they describe the Army environment for first-term Soldiers.

Results

The AES was administered to 107 Drill Sergeants and AIT instructors (E5-E7) during pilot testing. Only a few minor revisions were made to the original instrument based on initial data analysis and qualitative feedback from the NCOs. Prior to developing scale scores from the AES data, we examined whether there were any SMEs whose Likert-type ratings were highly inconsistent with the ratings of the other SMEs. To do so, we treated raters as "items" and computed item-total correlations between the ratings of each SME and the mean ratings of the other 106 raters. Results revealed five SMEs with near-zero or negative item-total correlations. As a result, the ratings of these individuals were excluded from further analysis.

[52] The environment-side supplies instruments were not designed to be administered to recruits in an operational setting. Rather, data from these measures will be used with data from the person-side needs measures (completed by recruits) to assess P-E fit. As such, with the exception of the Job Characteristics Survey (discussed later), none of the environment-side measures were administered beyond the pilot test data collections.

Table 13.2 displays descriptive statistics and reliability estimates for the AES Likert-type ratings. The internal consistency reliability estimates for the six AES scales were quite good (particularly for a 4-item scale), with all coefficients alpha but one (Realistic) being larger than .80. In addition, exploratory factor analysis (EFA)[53] indicated that six factors, representing the six interest dimensions, best described the NCO ratings.[54] Intraclass correlation coefficients (ICCs; McGraw & Wong, 1996) were used to estimate the consistency with which NCOs ordered the RIASEC dimensions. The resulting single- and *k*-rater[55] reliability estimates were .45 and .99, respectively.

Table 13.2. Descriptive Statistics and Reliability Estimates for AES and FAES Scale Scores

	AES					FAES					
Scale	Rank	*M*	*SD*	SE_M	α	Rank	*M*	*SD*	SE_M	α	*d*
Realistic	1	3.96	0.50	0.05	.71	2	4.17	0.34	0.14	.87	-0.55
Investigative	5	3.06	0.79	0.08	.88	5	2.63	0.82	0.33	.81	0.57
Artistic	6	2.58	0.87	0.09	.92	6	2.04	0.68	0.28	.78	0.74
Social	3	3.77	0.67	0.07	.83	1	4.25	0.44	0.18	.81	-0.86
Enterprising	4	3.62	0.71	0.07	.85	3	4.13	0.41	0.17	.56	-0.95
Conventional	2	3.79	0.63	0.06	.83	4	3.92	0.97	0.40	.83	-0.18

Note. $n = 102$ and 6 for AES and FAES ratings, respectively. Scale scores are based on items rated on a 5-point Likert-type scale. SE_M = standard error of the mean. *d* = standardized mean difference between AES and FAES ratings. These values were calculated by subtracting the FAES mean from the AES mean and dividing by the pooled *SD*. Results for the future-oriented version of the AES, the Future Army Environment Survey (FAES), are discussed in the next section of this chapter.

As for interrater agreement, the standard deviations (*SD*s) for the dimension ratings were 0.87 and below, and given the large number of NCOs who provided ratings, all of the standard errors of the mean were very low (i.e., .09 and below). One way to interpret the magnitude of the *SD*s is to compare them to *SD*s of (a) uniformly distributed (essentially random) ratings and (b) normally distributed ratings. The *SD* of uniformly distributed ratings for a 5-point scale is 1.41 and the *SD* of normally distributed ratings on such a scale is 1.15. All of the observed *SD*s are notably smaller than these values, which suggests that NCOs tended to agree about whether the current Army environment supports each type of interest (James, Demaree, & Wolf, 1984). Taken together, the quality of the AES ratings appears to be quite acceptable.

Figure 13.1 displays the RIASEC profile for the current Army work environment based on the AES scale scores. As expected, NCOs indicated that Realistic interests were supported most by the Army. Conventional, Social, and Enterprising interests were also rated relatively high, whereas Investigative and Artistic were clearly the two lowest rated dimensions. Although we had hoped to have more differentiation among the six dimensions (i.e., for assessing fit), repeated measures ANOVA revealed a statistically significant omnibus test across the six dimensions. Follow up tests (with a Bonferroni correction for multiple comparisons) indicated that most of the dimension means were significantly different.

[53] All EFA results reported in this chapter are based on a principal axis factor analysis with oblique rotation of factors.
[54] This finding should be interpreted cautiously given that the items comprising each AES scale were not independent (i.e., items were linked to a description of a specific RIASEC dimension).
[55] *k* = number of raters (i.e., 106). This coefficient estimates the reliability of AES ratings averaged across raters.

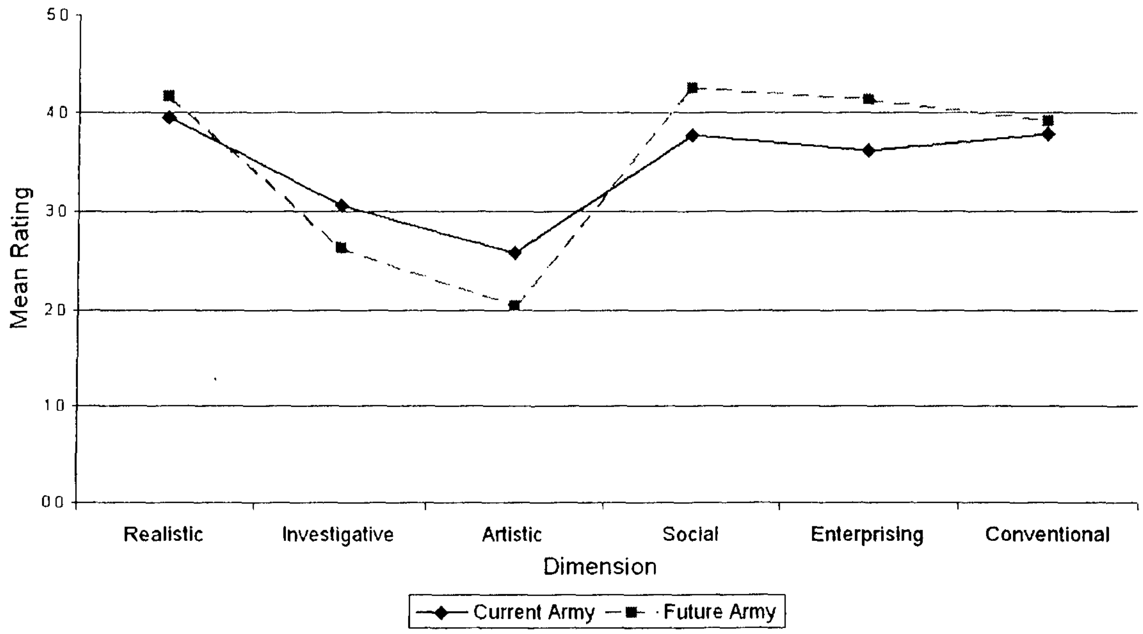

Figure 13.1. RIASEC interest profiles for the current and future Army.

Recall that NCOs were also asked to rank order the six dimensions in terms of how well they describe the Army work environment. Analyses of the AES ranking data revealed highly similar conclusions (to those based on the Likert-type rating data) about the extent to which the Army supports each RIASEC dimension. However, there are several reasons why we decided to not focus on the rank order results. First, fewer NCOs (n = 88) completed the rankings, as this rating exercise was added to the AES after initial data collection. Second, the *SD*s and standard errors of the dimension means were notably larger for the rankings than for the Likert-type ratings. Finally, we suspect that some respondents did not complete the ranking exercise as instructed, as there were many anomalous sets of ratings (e.g., NCOs who ranked Artistic interests as most supported by the Army). Taken together, only the Likert-type AES ratings will be used to assess P-E fit.

One of our main concerns in assessing fit with the general Army is that RIASEC profiles would differ by MOS. That is, we were concerned that the AES ratings the NCOs provided would reflect the work environment of their MOS and not the general Army. To examine this possibility, we compared AES profiles for five MOS (11B, 19D, 19K, 31U/74B, and 95B) for which we had a sufficient number of NCO ratings (n = 15 to 23 across the MOS) and found them to be highly similar across MOS. For instance, the zero-order correlations between the five AES profiles ranged from .91 to .99. This suggests that NCOs with different jobs have similar impressions of the general Army environment. There were, however, mean differences across MOS on individual AES scale scores. As we discuss below, the existence of such differences has implications for assessing P-E fit.

Future Army Environment Survey

Description of Measure

We also developed a version of the AES to determine RIASEC profile for the future Army. The Future Army Environment Survey (FAES) is identical to the AES except that respondents are asked to rate the extent to which they expect each interest to be supported by the future Army work environment.

Results

To obtain the RIASEC profile for the future Army, we asked the six members of the Select21 Subject Matter Expert Panel (SMEP) to complete the FAES. SMEP members reviewed the six descriptions of what the future Army environment is expected to be like (see Chapter 1), and then rated the extent to which they would expect the future Army to support each RIASEC dimension during Soldiers' first term of enlistment.

The results of FAES ratings are displayed in Table 13.2 along with the AES results.[56] Examination of the within-dimension *SD*s suggests that interrater agreement was higher for the FAES ratings than for the AES. Nonetheless, the standard errors of the mean were notably larger for the future ratings given the small number of raters. The internal consistency reliability estimates for the FAES ratings were acceptable (except for Enterprising). However, these reliability estimates should be interpreted cautiously given the small sample of raters. Finally, the single- and *k*-rater reliability estimates, .77 and .95 (respectively), indicate that the six SMEP members ordered the six dimensions in very similar ways.

The RIASEC profile for the future Army is very similar to the interests profile of the current Army (see Figure 13.1). Indeed, the zero-order correlation between the AES and FAES scale scores was .97. This suggests that the current and future Army might not differ in terms of the interests that are supported by the Army work environment. However, an alternative explanation for this result is that the SMEP ratings simply reflect perceptions of the current Army environment. Without knowledge of the future, it is difficult to determine the true level of similarity, although these results are generally consistent with our expectations. For example, we have little reason to believe that Artistic interests would be supported and that Realistic interests would not be supported as the Army transitions to the Future Force.

Despite strong relations between the current and future Army interest profiles, there were level differences on individual interest dimensions. For example, it appears that Realistic, Social, and Enterprising interests may be more supported in the future Army than in the current Army, whereas Investigative and Artistic interests may be less supported in the future. In other words, it appears that the future Army will provide greater support for the interests it currently supports and provide even less support for the interests it currently does not support.

---

[56] Given the small number of SMEs who provided FAES ratings, we did not use item-total correlations to identify SMEs with problematic ratings, as we did with the AES. However, a visual inspection of the FAES data did not reveal any aberrant raters.

As mentioned, dimension-level differences such as these have implications for how we combine person- and environment-side fit data to predict relevant criteria (see Appendix I). Specifically, the models we plan to use to combine these data are sensitive to dimension-level differences. Given this, even small differences between the current and future environment-side scores may alter conclusions we draw regarding the impact of a given scale on the criteria. Thus, we will likely need to consider both current and future environment-side interests (and values) data to assess fit during the concurrent validation.

Job Characteristics Survey

Description of Measure

As discussed, one of our main concerns about using Army-wide P-E fit interest measures for selection is that the RIASEC supplies profile would vary by MOS (i.e., there would be no "true" general Army profile). The comparison of AES results across NCOs from different MOS provides some evidence for the existence of an Army-wide interests profile. Nonetheless, given that vocational interests measurement is typically concerned with differentiating among vocations rather than among organizations, we also wanted to examine whether NCOs could differentiate between the interests their MOS supports and interests supported by the general Army work environment. If so, it would allow us to examine the relative (and perhaps incremental) validity of Army-wide and MOS-specific fit for predicting work attitudes and behavior.

To address the issue of Army-wide versus MOS-specific fit, we developed an MOS-specific interest measure called the Job Characteristics Survey (JCS). The JCS is identical to the AES and FAES except that respondents were asked to rate how well each RIASEC dimension describes their MOS, rather than the Army in general (e.g., "A first-term Soldier with realistic interests would be satisfied in your MOS.").

Results

The JCS was administered to 69 NCOs (E5-E7) during pilot testing. Analysis of the data did not reveal any problems with the instrument, and thus no changes were made. We also administered the JCS to an additional 71 NCOs (again, primarily E5-E7 Soldiers) during the criterion field test. The results described below are based on data combined from the two sets of SMEs.

Prior to creating scale scores for the JCS, we screened the data for problems. All 140 NCOs completed at least 90% of the JCS items, although the ratings of one NCO were excluded from further analysis due to a lack of variance in responses (i.e., the NCO rated all items a "5"). The remaining NCOs represented 16 different MOS. However, only six MOS (11B, 19D, 19K, 31U, 74B, and 96B) had what we considered a sufficient number of raters (i.e., at least 10) to create MOS-specific scale scores. Once these MOS were identified, we screened the individual MOS data sets for SMEs whose ratings were highly inconsistent with those of the other SMEs (using the same approach described earlier for the AES ratings). This process resulted in the ratings of six SMEs (across the six MOS) being excluded from further analysis. The JCS scale scores were based on ratings from 11 to 33 NCOs across the MOS (overall n = 107).

Table 13.3 displays descriptive statistics for the JCS scale scores. Analysis of the data revealed that one of the four scale items (the same item in each of the six scales) consistently detracted from scale reliability across MOS. As such, scale scores are based on only three items; however, the reliability estimates of the reduced scales were, in general, very good (median α = .87). Single- and k-rater reliability estimates, respectively, were also acceptable, ranging from .34 to .54 and from .91 to .96 across MOS.

Table 13.3. Descriptive Statistics for JCS Scale Scores by Represented MOS

	11B				19D				19K			
Scale	Rank	*M*	*SD*	SE_M	Rank	*M*	*SD*	SE_M	Rank	*M*	*SD*	SE_M
Realistic	1	3.71	0.71	0.12	1	3.64	0.97	0.26	1	3.92	0.58	0.16
Investigative	5	2.52	0.89	0.16	4	3.12	0.89	0.24	6	2.31	0.71	0.18
Artistic	6	2.18	1.02	0.18	6	1.86	0.86	0.23	5	2.47	0.66	0.17
Social	3	3.32	0.91	0.16	3	3.19	0.75	0.2	2	3.62	0.56	0.14
Enterprising	2	3.47	1.00	0.17	2	3.55	0.58	0.15	4	3.56	0.56	0.14
Conventional	4	3.17	0.93	0.16	5	2.88	1.29	0.35	3	3.60	0.81	0.21
	31U				74B				96B			
Realistic	2	3.82	0.70	0.21	4	3.62	1.28	0.34	4	3.02	0.93	0.21
Investigative	5	2.52	1.21	0.36	5	3.50	0.85	0.23	1	3.93	0.77	0.17
Artistic	6	2.09	0.94	0.28	6	2.57	1.13	0.3	6	2.47	0.78	0.17
Social	3	3.21	0.79	0.24	2	4.00	0.65	0.17	5	2.83	0.76	0.17
Enterprising	4	2.90	1.05	0.33	3	3.76	0.62	0.17	2	3.90	0.33	0.07
Conventional	1	3.91	0.62	0.19	1	4.36	0.44	0.12	3	3.47	0.69	0.15

Note. n_{11B} = 33, n_{19D} = 14, n_{19K} = 15, n_{31U} = 11, n_{74B} = 14, and n_{96B} = 20. SE_M = standard error of the mean. Scale scores are based on items rated on a 5-point Likert-type scale.

Figure 13.2 shows the RIASEC profiles for the represented MOS. Clearly, there are some profile differences among these MOS. For example, according to SMEs, Investigative interests are supported to a much greater extent in the 96B work environment than in the 19K environment. To assess the magnitude of these differences, we computed zero-order correlations between the profiles. Correlations ranged from -.03 between 19K and 96B to .95 between 11B and 19K (median r = .65). These results provide evidence that RIASEC profiles differ across at least some MOS.

We also compared the JCS MOS-specific scale scores to AES scale scores to see whether differences emerged between MOS and Army in general interest profiles. Zero-order correlations between the general Army profile and the 11B, 19D, 19K, 31U, 74B, and 96B profiles (respectively) were .96, .81, .93, .92, .84, and .22. Thus, with the exception of 96B, the MOS interest profiles were very similar to the Army profile. Nonetheless, there were some rather notable JCS-AES mean differences on individual interest dimensions. For example, despite the very strong relationship between the 11B and general Army profiles, the AES Conventional scale score (M = 3.79) was significantly higher (p < .05) than the corresponding 11B scale score (M = 3.17).

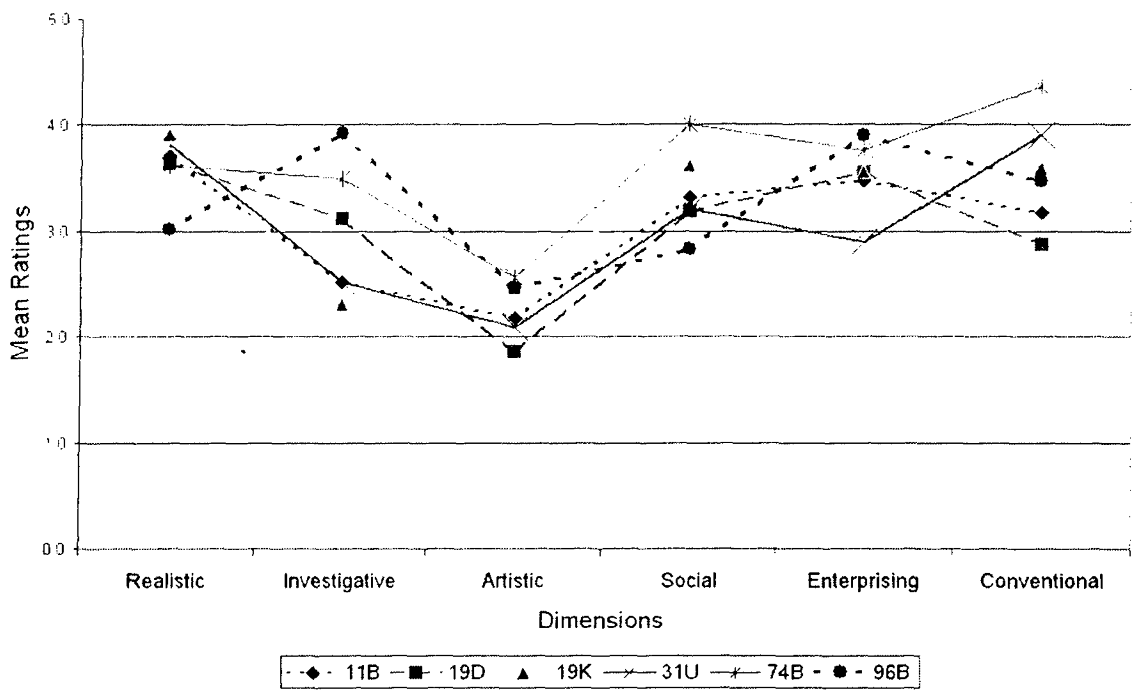

Figure 13.2. RIASEC interest profiles for represented MOS.

Discussion

In summary, the three environment-side vocational interest measures (i.e., the AES, FAES, and JCS) have provided valuable data for measuring P-E fit. Although we are generally satisfied with the psychometric quality of these measures, the results do raise a few issues that will likely affect use of these measures during both the concurrent validation and attrition database analyses (see Putka, 2004). For example, given the similarity of current and future Army interest profiles, it may be redundant to consider both profiles in future research. Nevertheless, given some of the differences between the current and expected future Army on individual interest dimensions, and the fact that the method for using such data to predict criteria is sensitive to differences at the interest-level (see Appendix I), the question of redundancy will most likely be addressed empirically.

Another issue is the relatively large standard errors around the FAES scale score. This was partly due to the small number of SMEP members who provided ratings, but also to the fact that SMEP members were selected to maximize coverage of knowledge regarding the future Army. Specifically, each SMEP member had specialized knowledge about one or more aspects of the future Army, yet no single member was an authority in all areas. This heterogeneity may help account for some of the variability in their FAES ratings. Unfortunately, very few people in the Army can provide insight as to what the future work environment will be like for first-term enlisted Soldiers, and thus no additional future ratings will be collected. The standard errors for the JCS scale means are also rather high (e.g., SE_M = .19 to .36 across dimensions for 31U). However, unlike the FAES, we are able to collect additional JCS data for the represented MOS prior to the concurrent validation to increase the stability of the JCS scale scores.

Person-Side Needs Measures

Work Preferences Survey

Description of Measure

The AES, FAES, and JCS are supplies measures for assessing needs-supplies fit. We also have developed two needs measures that in operation would be used to generate interest profiles for potential recruits. These profiles could, in turn, be compared to one or more supplies profiles to assess fit (e.g., with the current Army) for selection. Conversely, these needs measures could be administered post-selection (e.g., to new recruits during accession) to help inform placement decisions and/or identify recruits whose interests do not match well with those supported by the Army work environment.

The first needs measure is the Work Preferences Survey (WPS). The 75-item WPS contains three types of items that measure interests in work activities (e.g., "A job that requires me to teach others"), work environments (e.g., "A job that requires me to work outdoors"), and learning opportunities (e.g., "A job in which I can learn how to lead others") related to the RIASEC dimensions. Respondents are asked to rate each item on a Likert-type scale with anchors that range from *extremely unimportant to have in my ideal job* (1) to *extremely unimportant to have in my ideal job* (5). Item development was based on a thorough review of existing interest inventories and source materials from the vocational interest literature.

Pilot Test Results

Initial versions of the WPS were administered to over 400 new recruits during pilot testing. These data were subjected to a variety of item- and scale-level analyses (e.g., internal consistency reliability analysis, EFA). We also examined relations between the WPS and an adapted version of an existing interest measure, the Interest Finder Questionnaire (IFQ; discussed later). In addition, we revised and eliminated items based on qualitative feedback from respondents (e.g., items that did not make sense). In general, the WPS dimensions appeared to work quite well. We did, however, revise or delete about 10% of the original items on the basis of the pilot test data. We also added several new items.

Faking Research Results

The WPS was administered during the Select21 faking research to assess its susceptibility to response distortion. A mixed between-within subjects design was used whereby all new recruits who completed the WPS (n = 196) first did so under honest instructions. Then, about half of the recruits (n = 96) were instructed to present themselves in the best possible light (i.e., the "fake max" condition) and half (n = 100) were instructed to fake but try to avoid being detected (i.e., the "fake avoid detection" condition).[57]

The key results for the WPS are displayed in Table 13.4. Instructing recruits to fake had little influence on the *SD*s or internal consistency estimates for the WPS dimensions. However, as expected, instructions to fake resulted in significantly higher mean ratings on all dimensions except Artistic. Conventional was the most inflated dimension in both the fake max and avoid

[57] The specific instructions for these three conditions are provided in Appendix F.

detection conditions (d = 1.22 and 0.73, respectively), whereas Artistic was the least inflated (d = 0.10 and -0.29). Thus, it appears that many recruits attempted to respond in line with an Army profile, rather than inflating their responses on all dimensions. For example, compared to their honest responses, recruits gave relatively higher ratings to Realistic items and relatively lower ratings to Artistic items. That said, the RIASEC profiles based on the faking data do not precisely match the profile for the current Army based on the AES results. This suggests that even when motivated, some individuals may not be able to distort their responses to increase their fit score. It is also interesting that even in the fake max condition, the dimension means did not approach the upper limit of the 5.0 scale. For instance, the mean rating for Realistic in the fake max condition was only 4.14, which was somewhat surprising given that Realistic items would seem to be most clearly relevant to the Army work environment.

We also computed zero-order correlations between honest and faking condition dimension scores (see Table 13.4). These relatively low coefficients (median r = .27) suggest that instructing recruits to fake resulted in a very different ordering of respondents (in terms of each interest dimension) relative to when they completed the WPS honestly. The low correlations provide additional evidence for individual differences in faking behavior. That is, if everyone faked to the same extent, then relations between honest and faking scores should be quite high.

Next, we calculated item-level honest-fake differences to see whether there were any particularly problematic WPS items, but none were found. Given this, we did not use the faking results to revise the WPS. We did, however, use this opportunity to further examine the psychometric characteristics of the WPS items and scales from the honest condition. Based on these results, we eliminated 2 items, revised 5 items, and added 10 new items. Several of the new items were based on an examination of 1- to 6-month attrition data for Soldiers who completed the WPS during pilot testing. For example, items designed to measure interests in physical-related work activities (which are a facet of the Realistic dimension) were particularly predictive of attrition. As such, we developed a few additional items that could be used to create a physical subscale.

To summarize, although individuals can clearly inflate their responses to the WPS when instructed, not all recruits in this sample inflated to the same degree and or with the same level of "accuracy" (i.e., in relation to the actual Army profile). Given this, it remains to be seen whether and how response distortion will affect the assessment of P-E fit.

Field Test Results

Sample. The WPS was administered to 693 new recruits during the predictor field test. Of these, 59 recruits failed to complete at least 90% of the WPS items. Responses from two additional recruits were excluded from further analysis because test administrators flagged these individuals as having questionable WPS data. Thus, the final analysis sample comprised 632 cases, or 91.2% of the initial sample.

Data analysis. We began by examining item-level statistics for each WPS scale, including means, *SD*s, and item-deleted reliability statistics, to identify problematic items. We also used EFA to assess the dimensionality of items comprising each scale. We then computed descriptive statistics and intercorrelations for the revised scales. Next, confirmatory factor analysis (CFA) was used to assess the fit of the a priori 6-factor model. The CFA was conducted with LISREL 8.3 (Jöreskog & Sörbom, 1996) on the covariance matrices using maximum

Table 13.4. Descriptive Statistics, Reliability Estimates, and Fake-Honest Differences for the WPS and IFQ Scale Scores

		Honest			Fake Max			Avoid Detection			FM-H		AD-H		FM-AD
Measure/Scale	Items	*M*	*SD*	α	*M*	*SD*	α	*M*	*SD*	α	*d*	*r*	*d*	*r*	*d*
WPS															
Realistic	10	3.29	0.86	.89	4.14	0.67	.84	3.91	0.82	.89	0.99	.30	0.72	.31	0.31
Investigative	10	3.57	0.70	.86	4.17	0.76	.88	3.89	0.63	.78	0.86	.18	0.46	.30	0.40
Artistic	11	3.03	0.79	.87	3.11	1.04	.92	2.80	0.91	.90	0.10	.28	-0.29	.27	0.32
Social	9	3.55	0.72	.86	4.29	0.67	.84	3.97	0.65	.78	1.03	.21	0.58	.29	0.48
Enterprising	13	3.34	0.64	.84	3.87	0.72	.84	3.69	0.57	.75	0.83	.34	0.55	.23	0.28
Conventional	12	3.27	0.67	.85	4.09	0.64	.84	3.76	0.58	.74	1.22	.23	0.73	.32	0.54
IFQ															
Realistic	15	1.99	0.54	.90	2.51	0.49	.91	2.36	0.50	.89	0.79	.31	0.85	.33	0.30
Investigative	25	2.03	0.48	.92	2.53	0.45	.94	2.37	0.41	.90	1.03	.30	0.70	.24	0.37
Artistic	15	1.88	0.50	.88	1.97	0.57	.92	1.69	0.47	.87	0.29	.22	-0.49	.24	0.54
Social	15	1.99	0.48	.87	2.57	0.38	.83	2.38	0.36	.74	1.24	.18	0.75	.18	0.51
Enterprising	15	2.07	0.50	.88	2.37	0.42	.83	2.22	0.38	.75	0.59	.22	0.28	.19	0.38
Conventional	15	1.66	0.52	.90	2.52	0.43	.87	2.30	0.46	.86	1.75	.12	1.15	.28	0.49

Note. *n* for Honest, Fake Max, and Avoid Detection = 196, 96, and 100, respectively. FM = Fake Max condition. H = Honest condition. AD = Avoid Detection condition. WPS and IFQ scale scores are based on 5- and 3-point Likert-type items, respectively. *d* = standardized mean difference, which was calculated by subtracting the honest mean from the faking mean and dividing by the *SD* from the honest condition. *r* = zero-order correlation between honest and faking dimension scores. FM-H and AD-H *d* and *r* statistics are based on within-subjects data.

likelihood estimation. We also examined the subgroup effect sizes (with regard to gender and race/ethnicity) associated with the WPS scale scores. Finally, we report descriptive statistics and subgroup effect sizes for fit indices based on scale scores from the WPS and environment-side supplies measures (i.e., the AES and FAES).[58] Results of these analyses are described in turn.

Scale refinement. Analysis of the WPS scales revealed only a few problematic items. Specifically, four items were eliminated due to low item-scale correlations and/or because they correlated more highly with another WPS scale than the intended scale. EFA of the items within each scale indicated that all scales were at least somewhat multifaceted. For example, the Realistic items loaded on two related, yet distinct, factors, representing interests in physical and mechanical work activities. Likewise, the Social items comprised two main factors—one dealing with interests in working with others and the other with interests in helping others.

EFA of all the WPS items (minus the four problematic ones) revealed that five rather than six factors appeared to best describe the data. Social and Enterprising items loaded on the same factor instead of on separate factors. The main reason for this overlap is that both scales include items about working with people. The theoretical difference is that Enterprising individuals are interested in leading and directing others, whereas individuals with Social interests tend to like teaching and helping others. Although this lack of differentiation is somewhat of a concern, a strong association between Social and Enterprising interests has been reported elsewhere in the literature (e.g., Project A; J. P. Campbell & Knapp, 2001). Also of note is that there continues to be some overlap between certain facets of Investigative and Conventional interests. Specifically, Conventional items that assess attention to detail and being organized tended to load on the Investigative factor rather than on their intended factor. Again, although challenging from a measurement perspective, this association is not entirely surprising given some of the conceptual similarities between these two interest dimensions.

Descriptive statistics and reliability estimates. Table 13.5 shows descriptive statistics and reliability estimates for the revised WPS scale scores. Examination of the scale means suggests that recruits in this sample have a wide range of occupational interests, including those the general Army work environment does not tend to support (e.g., Artistic interests). The internal consistency estimates of the WPS scales were very good, with all reliability estimates being .84 or higher.

This table also displays intercorrelations among the WPS scale scores. With the exception of Realistic, most WPS scales were at least moderately correlated (median $r = .42$). As expected given the EFA results, Social and Enterprising scale scores exhibited the strongest correlation ($r = .70$). Interestingly, Investigative correlated notably higher with Enterprising ($r = .67$), Conventional ($r = .63$), and Social ($r = .62$) than in prior data collections (e.g., rs < .45 in the faking research sample). As discussed, the strong relation between Investigative and Conventional scores was due primarily to the association between Conventional attention to detail and organization items and Investigative scale scores. Conversely, there is no immediate explanation for the high relations between Investigative and the other two WPS scales (e.g., there were no subsets of logically related items that demonstrated large cross-scale correlations).

[58] We did not compute MOS fit indices for any of the person-side interest measures due to relatively small MOS-specific sample sizes on which the JCS scale scores are based. However, given our plans to collect additional JCS data, we will assess MOS fit during the concurrent validation and attrition database analyses.

Table 13.5. Descriptive Statistics, Reliability Estimates, and Intercorrelations for WPS and IFQ Scale Scores

Measure/Scale	Items	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12
WPS															
1 Realistic	12	3.40	0.83	(.91)											
2 Investigative	12	3.29	0.71	.06	(.88)										
3 Artistic	12	2.77	0.76	.01	**.43**	(.88)									
4 Social	12	3.45	0.71	.00	**.62**	**.42**	(.84)								
5 Enterprising	10	3.47	0.65	**.08**	**.67**	**.34**	**.70**	(.84)							
6 Conventional	13	3.17	0.67	-.02	**.63**	**.25**	**.58**	**.58**	(.85)						
IFQ															
7 Realistic	14	2.06	0.57	**.56**	-.01	.05	-.07	-.05	.03	(.90)					
8 Investigative	21	2.09	0.56	.04	**.51**	**.32**	**.26**	**.28**	**.21**	**.24**	(.93)				
9 Artistic	15	1.96	0.52	-.05	.24	**.56**	**.28**	**.19**	**.12**	**.19**	**.41**	(.88)			
10 Social	13	2.04	0.50	**-.12**	**.36**	**.30**	**.53**	**.39**	**.38**	**.13**	**.47**	**.53**	(.85)		
11 Enterprising	14	2.15	0.49	-.07	**.35**	**.28**	**.33**	**.46**	**.28**	**.15**	**.47**	**.41**	**.56**	(.86)	
12 Conventional	15	1.77	0.55	**-.17**	**.27**	**.14**	**.30**	**.26**	**.59**	**.19**	**.25**	**.23**	**.51**	**.42**	(.91)

Note. $n = 632$ and 658 for the WPS and IFQ, respectively. Correlations between WPS and IFQ scores are based on $n = 606$. Scale scores for the WPS and IFQ are based on 5- and 3-point Likert-type items, respectively. Internal consistency reliability estimates (alpha) are shown along the diagonal in parentheses. Convergent correlations are in boxes. Statistically significant correlations are bolded, $p < .05$ (two-tailed).

Model fit. We then assessed the fit of the WPS measurement model using CFA. Given the high correlations between several of the WPS scale scores, it was not surprising that the CFA did not reveal a very good fit for the 6-factor model (χ^2 (2,399) = 9381.97, $p < .01$; GFI = .60; CFI = .68; RMSEA = .090).[59] Although all items loaded significantly on their intended factor, examination of the residuals indicated that allowing various items to cross-load on one or more additional factors would notably improve model fit. The poor fit was also due, in part, to the multifaceted nature of the WPS scales. Specifically, certain sets of items within some of the WPS scales (e.g., physical-related items within the Realistic scale) share more variance than that accounted for by the overall interest factor. Indeed, results indicated that including error covariances between such items would significantly enhance the fit of the model. The large number of indicators per latent factor (i.e., 10 to 13),[60] and the modest ratio of sample size to indicators (i.e., about 9:1) also might have contributed to the general lack of support for the measurement model.

Subgroup differences. Next, we examined subgroup differences associated with the WPS scale scores. Results of these analyses appear in Tables 13.6 and 13.7 for gender and race/ethnicity, respectively. We found several moderate and statistically significant gender differences across the WPS scales. Specifically, male recruits had higher Realistic interests, whereas female recruits had higher Artistic, Social, and Conventional interests. These differences are generally consistent with those reported in prior research (e.g., J. P. Campbell & Knapp, 2001). There were also significant subgroup differences with regard to race. The largest effect was that Black and Hispanic recruits had higher Conventional interests than White recruits. Minority recruits also had significantly higher Social and Enterprising interests. In contrast, Whites had higher Realistic interests than both Blacks and Hispanics.

Fit indices. In the final set of WPS analyses, we calculated two common fit indices (D^2 and Pearson's r) to assess the fit between recruits' interests (measured by the WPS) and the interests supported by the current and future Army work environment (measured by the AES and FAES, respectively).[61] The D^2 index is calculated by summing the squared differences between recruits' mean scores on each WPS scale and the corresponding mean scores from the supplies measure. Thus, smaller D^2 values indicate a better needs-supplies fit. As a point-of-reference, if WPS scale scores for a given recruit differed from each of the corresponding AES scores by .50, 1.0, and 2.0 scale points, the resulting D^2 values would be 1.5, 6.0, and 24.0, respectively. The r index is a simple zero-order correlation between recruits' interest profile and the environment-side profile. As such, larger values indicate a good fit.

---

[59] χ^2 = chi-square statistic. GFI = goodness-of-fit-index. CFI = comparative fit index. RMSEA = root-mean-square error of approximation.

[60] A potential problem with using single-item indicators in CFA is that individual items often possess a notable amount of unique variance (i.e., variance unaccounted for by the factor of interest). Thus, including numerous single-item indicators can limit the amount of variance in the indicators that can be accounted for by the factor(s), and thereby result in lower estimates of model fit.

[61] Fit indices for all of the measures discussed in this chapter were computed primarily for descriptive purposes, not necessarily for operational use. Although they are useful for describing the similarity between recruits' vocational interests (work values) and those supported by the Army work environment, fit indices such as this can be problematic when used for prediction (see Appendix I for details). Our plans for combining person- and environment-side data to assess fit during the concurrent validation and for potential operational use are described in Appendix I. Nevertheless, we will evaluate the potential utility of these fit indices (i.e., D^2 and Pearson's r) within the Select21 attrition database.

Table 13.6. WPS and IFQ Scale Scores by Gender

		Male		Female	
Measure/Scale	d_{FM}	*M*	*SD*	*M*	*SD*
WPS					
Realistic	**-0.60**	3.54	0.79	3.07	0.80
Investigative	0.15	3.26	0.74	3.37	0.64
Artistic	**0.18**	2.73	0.78	2.87	0.71
Social	**0.62**	3.24	0.71	3.69	0.59
Enterprising	0.13	3.32	0.67	3.41	0.58
Conventional	**0.59**	3.05	0.65	3.43	0.61
IFQ					
Realistic	**-0.41**	2.13	0.56	1.90	0.56
Investigative	-0.11	2.11	0.57	2.05	0.52
Artistic	**0.42**	1.90	0.51	2.12	0.51
Social	**0.79**	1.92	0.50	2.31	0.39
Enterprising	-0.08	2.16	0.51	2.12	0.45
Conventional	**0.74**	1.66	0.51	2.03	0.55

Note. For the WPS, $n_{Male} = 443$, $n_{Female} = 185$. For the IFQ, $n_{Male} = 462$, $n_{Female} = 191$. WPS and IFQ scale scores are based on 5-and 3-point Likert-type items, respectively. d_{FM} = Effect size for Female-Male mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., Males) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

Table 13.7. WPS and IFQ Scale Scores by Race/Ethnic Group

			White		Black		White Non-Hispanic		Hispanic	
Measure/Scale	d_{BW}	d_{HW}	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
WPS										
Realistic	**-0.51**	**-0.53**	3.50	0.80	3.09	0.87	3.55	0.77	3.14	0.84
Investigative	0.15	**0.29**	3.26	0.72	3.37	0.74	3.23	0.73	3.44	0.58
Artistic	0.18	0.01	2.72	0.76	2.86	0.68	2.73	0.77	2.74	0.65
Social	**0.47**	**0.36**	3.30	0.69	3.63	0.72	3.27	0.69	3.52	0.55
Enterprising	**0.35**	**0.33**	3.29	0.62	3.51	0.67	3.27	0.62	3.47	0.62
Conventional	**0.64**	**0.62**	3.07	0.64	3.48	0.68	3.02	0.63	3.41	0.53
IFQ										
Realistic	**-0.39**	**-0.40**	2.12	0.57	1.90	0.49	2.15	0.56	1.92	0.55
Investigative	-0.12	0.14	2.10	0.57	2.03	0.57	2.08	0.58	2.16	0.51
Artistic	0.14	0.02	1.93	0.52	2.00	0.54	1.93	0.52	1.94	0.52
Social	**0.25**	0.19	2.01	0.52	2.15	0.45	2.00	0.53	2.10	0.45
Enterprising	**0.29**	0.20	2.11	0.50	2.26	0.47	2.10	0.50	2.20	0.47
Conventional	**0.57**	**0.56**	1.70	0.54	2.01	0.57	1.66	0.54	1.97	0.51

Note. For the WPS, $n_{White} = 418$. $n_{Black} = 96$. $n_{White\ Non\text{-}Hispanic} = 367$. $n_{Hispanic} = 78$. For the IFQ, $n_{White} = 427$. $n_{Black} = 104$. $n_{White\ Non\text{-}Hispanic} = 371$. $n_{Hispanic} = 83$. WPS and IFQ scale scores are based on 5-and 3-point Likert-type items, respectively. d_{BW} = Effect size for Black-White mean difference. d_{HW} = Effect size for Hispanic-White Non-Hispanic mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., White) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

Descriptive statistics and intercorrelations for these fit indices are shown in Table 13.8. Several findings are noteworthy. First, given the similarity of AES and FAES interest profiles (discussed earlier), it is not surprising that fit indices based on these two measures were highly correlated (e.g., .98 for the *r* indices). On the other hand, the two types of fit indices (D^2 and *r*) were only moderately related. The difference between these indices is that D^2 reflects differences between profiles in terms of elevation (i.e., differences in profile means), scatter (i.e., differences in profile *SD*s), and shape (i.e., differences in the ordering of profile elements), whereas *r* reflects differences only in terms of shape (Cronbach & Gleser, 1953). Thus, the moderate correlations between the two fit indices indicate that there were differences between WPS and AES/FAES profiles with regard to elevation and/or scatter. Lastly, results revealed wide variation in fit indices across recruits in this sample (e.g., WPS-AES r = -.98 to .98).

Table 13.8. Descriptive Statistics and Intercorrelations for WPS and IFQ Fit Index Scores

Fit Index	*M*	*SD*	1	2	3	4	5	6	7	8
WPS										
1 Current Army D^2	4.15	3.66	–							
2 Current Army *r*	0.31	0.45	-.33	–						
3 Future Army D^2	6.64	4.23	.97	-.48	–					
4 Future Army *r*	0.31	0.42	-.33	.98	-.50	–				
IFQ										
5 Current Army D^2	3.73	2.27	.35	-.11	.35	-.12	–			
6 Current Army *r*	-0.01	0.44	-.17	.51	-.25	.49	-.32	–		
7 Future Army D^2	5.01	2.36	.35	-.21	.38	-.22	.97	-.47	–	
8 Future Army *r*	0.06	0.41	-.17	.51	-.26	.51	-.30	.96	-.48	–

Note. n = 632 and 658 for the WPS and IFQ, respectively. Correlations between WPS and IFQ fit indices are based on n = 606. Fit indices for current and future Army are based AES and FAES profile scores, respectively. All correlations are statistically significant, $p < .05$ (two-tailed).

We also calculated subgroup effects sizes for the fit indices (see Tables 13.9 and 13.10). Results showed some statistically significant mean differences, all of which favored recruits from majority groups (i.e., males and Whites). Nevertheless, the effect sizes associated with these differences were small in magnitude (Cohen, 1992).

Discussion

The overall results of the data analysis suggest promise for the WPS. For example, scale scores were reliable and produced sufficient variation among respondents. We also found some evidence for the construct-related validity of the WPS scales in relation to IFQ scale scores (discussed later).

Nonetheless, we do have some concerns about the WPS. First, despite efforts to reduce overlap between dimensions, correlations between some of the WPS scale scores (e.g., Social and Enterprising) remain higher than we would like. Given this, we will further examine the WPS to ensure that items comprising each scale appropriately sample the content domain and attempt to eliminate any unnecessary overlap between scales. A secondary concern is some of the WPS subgroup differences that emerged during the field test. Although most of the score differences favored individuals from the minority groups, male and White recruits had notably higher Realistic scores than minority group members. Such differences could result in adverse

impact and/or predictive bias if the WPS were used for selection. It is important to note, however, that these differences did not extend to the WPS fit index scores. In any case, we will be mindful of subgroup differences (and their effects) during the concurrent validation.

Table 13.9. WPS and IFQ Fit Index Scores by Gender

		Male		Female	
Measure/Fit Index	d_{FM}	*M*	*SD*	*M*	*SD*
WPS					
Current Army D^2	-0.13	4.31	3.93	3.78	2.94
Current Army *r*	-0.14	0.33	0.47	0.26	0.41
Future Army D^2	**-0.15**	6.85	4.54	6.15	3.37
Future Army *r*	-0.04	0.31	0.44	0.29	0.38
IFQ					
Current Army D^2	**-0.30**	3.95	2.35	3.25	1.99
Current Army *r*	-0.09	0.00	0.43	-0.04	0.45
Future Army D^2	**-0.23**	5.19	2.45	4.63	2.09
Future Army *r*	-0.16	0.08	0.41	0.02	0.41

Note. For the WPS, $n_{Male} = 443$, $n_{Female} = 185$. For the IFQ, $n_{Male} = 462$, $n_{Female} = 191$. Fit indices for current and future Army are based AES and FAES profile scores, respectively. d_{FM} = Effect size for Female-Male mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., Males) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

Table 13.10. WPS and IFQ Fit Index Scores by Race/Ethnic Group

			White		Black		White Non-Hispanic		Hispanic	
Measure/Fit Index	d_{BW}	d_{HW}	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
WPS										
Current Army D^2	0.03	**-0.21**	4.11	3.22	4.21	4.89	4.17	3.33	3.46	2.14
Current Army *r*	-0.03	-0.02	0.33	0.44	0.32	0.43	0.33	0.44	0.33	0.42
Future Army D^2	-0.03	**-0.23**	6.62	3.77	6.50	5.52	6.70	3.88	5.82	2.73
Future Army *r*	0.07	0.05	0.32	0.40	0.35	0.41	0.32	0.41	0.34	0.40
IFQ										
Current Army D^2	**-0.23**	**-0.26**	3.86	2.30	3.33	2.19	3.94	2.37	3.31	1.84
Current Army *r*	0.11	0.00	0.00	0.42	0.05	0.48	0.00	0.42	0.00	0.44
Future Army D^2	-0.21	-0.20	5.11	2.37	4.61	2.27	5.17	2.44	4.68	2.01
Future Army *r*	0.07	-0.11	0.08	0.39	0.11	0.44	0.09	0.39	0.04	0.40

Note. For the WPS, $n_{White} = 418$. $n_{Black} = 96$. $n_{White\ Non\text{-}Hispanic} = 367$. $n_{Hispanic} = 78$. For the IFQ, $n_{White} = 427$. $n_{Black} = 104$. $n_{White\ Non\text{-}Hispanic} = 371$. $n_{Hispanic} = 83$. Fit indices for current and future Army are based AES and FAES profile scores, respectively. d_{BW} = Effect size for Black-White mean difference. d_{HW} = Effect size for Hispanic-White Non-Hispanic mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., White) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

Lastly, we recommend computing WPS fit indices based on only the current Army interests profile (derived from the AES data) in the concurrent validation and for the attrition database analyses. This recommendation is based on several findings, including the very strong correlation between current and future Army interest profiles (and between the fit indices based

on those profiles) and the higher level of rater agreement for the current than for the future Army profile. Nevertheless, given some of the mean differences between AES and FAES scores on specific interest dimensions and the sensitivity of our planned method for combining person- and environment-side interest data to such differences, we will consider both current and future supplies data when modeling WPS-criterion relations in subsequent research.[62]

Interest Finder Questionnaire

Description of Measure

The IFQ is the second vocational interest needs measure. The IFQ is an adapted version of the Interest-Finder developed by Defense Manpower Data Center (DMDC) for vocational counseling (Wall, Wise, & Baker, 1996). In the IFQ, respondents are asked to rate their interest in 98 work activities (that reflect the RIASEC dimensions) by indicating whether they *like* (1), *don't know* (2), or *dislike* (3) each activity. The initial purpose of the IFQ was to serve as a marker measure with which to assess construct-related validity of the WPS. However, the two instruments measure interests in slightly different ways, and at this point it is unclear which method might be more effective on factors such as resistance to response distortion and construct- and criterion-related validity. Given this, we have continued to evaluate and refine the IFQ for potential operational use.

Pilot Test Results

We began by using the entire Interest-Finder instrument, which consists of 240 items that assess interest in activities (e.g., "host social events"), training opportunities (e.g., "how to start your own business"), and occupations (e.g., "district attorney"). However, during pilot testing we found that the activities items demonstrated better psychometric characteristics (e.g., factor structure, internal consistency) than the training and occupation items.[63] Because of this and instrument length, we eliminated the training and occupations items. However, we retained some of the training items that performed well by transforming them into activities items. We also eliminated and revised a few items that pertained to outdated activities that many Soldiers indicated were unfamiliar to them (e.g., "Use a battery tester"). In addition, we created several new Investigative items to broaden the focus of that dimension beyond the "hard" sciences. These revisions resulted in a 100-item instrument that was used in the faking research discussed below.

Faking Research Results

The same sample of new recruits who completed the WPS also completed the IFQ under the same honest and faking instructions. The dimension level results are shown in Table 13.4. The overall pattern of results was fairly consistent with those for the WPS. For example, instructing recruits to fake had little influence on the *SD*s or internal consistency estimates of the RIASEC scales (although the alphas in the avoid detection condition were somewhat lower). However, instructions to fake resulted in significantly higher mean ratings on all dimensions except Artistic. Consistent with the WPS results, Conventional was the most inflated dimension

[62] These recommendations also apply to the other two interest predictors (i.e., IFQ and Pre-Service Expectations Survey).
[63] DMDC has found similar results and has developed a new interest inventory that comprises only activities items.

in both the fake max and avoid detection conditions (*d* = 1.75 and 1.15, respectively), whereas Artistic was the least inflated (*d* = 0.29 and -0.49). Also consistent with the WPS results is that many recruits attempted to fake in line with an Army profile (e.g., by deemphasizing artistic interests), rather than uniformly inflating their responses on all dimensions. Lastly, the zero-order correlations between honest and faking condition scores were quite low (median *r* = .23), which provides evidence for individual differences in faking.

We examined honest-fake differences at the item level, but did not eliminate or revise any IFQ items based on those results. We also looked at the psychometric characteristics of IFQ items and scales using data from the honest condition. The overall validity and reliability evidence for the IFQ dimensions was quite good. For example, factor analysis revealed six fairly "clean" factors that represented the intended RIASEC dimensions. We did, however, revise eight items and eliminate two items based on the analysis results. This resulted in the 98-item instrument used in the field test.

Field Test Results

Sample. The IFQ was administered to 672 new recruits during the predictor field test. Of these, data from 11 recruits were excluded from analysis because they failed to complete 90% of items and/or lacked variance in their responses. Data from an additional respondent was eliminated based on comments in the field test problems log. Thus, the analysis sample comprised data from 658 recruits, or 97.9% of the initial sample.

Scale refinement. The same analyses we performed on the WPS field test data also were used to assess the IFQ data. We began by examining item-level statistics for each IFQ scale. Six problematic items were identified. These items were excluded from subsequent analyses. As with the WPS, EFA of the items within each IFQ scale suggested that all scales comprised multiple facets. Artistic items, for example, loaded on three related factors that reflected interests in writing, music, and sculpting and decorating. In addition, the Conventional items were best described by two factors, which measured interests in administrative activities and interests in accounting and finance. However, unlike the WPS, EFA of all the IFQ items suggested that six factors (representing the six interest dimensions) best described the data. In fact, very few items had notable cross-loadings (i.e., $\geq$ .30) on other factors. One possible reason why the structure of the IFQ data appears to be better represent the RIASEC dimensions is that it only assesses interest in work activities. In contrast, the WPS comprises multiple item types, which measure interest-related activities, environments, and learning opportunities.

Descriptive statistics and reliability estimates. Table 13.5 presents descriptive statistics and reliability estimates, and intercorrelations for the revised IFQ scale scores. As with the WPS, the scale means were rather undifferentiated, which suggests that recruits in this sample have a variety of vocational interests. The reliability estimates for the IFQ scales were very good, ranging from .85 (Social) to .93 (Investigative) across dimensions. As for relations among the scale scores, the overall magnitude of the correlations (median *r* = .41) was almost identical to that of the WPS (median *r* = .42). Also consistent with the WPS results was that Realistic was the most distinct IFQ scale and Social and Enterprising were the most correlated (*r* = .56). However, unlike the WPS, the IFQ scale correlations were, in general, smaller than those found in prior data collections (e.g., in the faking research data).

We also examined relations between scales from the WPS and IFQ using multitrait-multimethod analysis (D. T. Campbell & Fiske, 1959). As Table 13.5 shows, all convergent (i.e., same dimension, different instrument) correlations were statistically significant, and the median convergent correlation (.56) was significantly larger ($p < .05$) than the median discriminant (i.e., different dimension, same instrument) correlation (.42). This pattern of relations is very similar (yet slightly better in terms of convergent and discriminant evidence) to relations between these two instruments found in the faking research data. Thus, although the WPS and IFQ measure occupational interests in somewhat different ways, these results provide some evidence that they are measuring similar constructs.

Model fit. CFA was used to assess the fit of the measurement model within the IFQ data. Results of this analysis were somewhat mixed (χ^2 (4,079) = 10,498.58, $p < .01$; GFI = .70; CFI = .75; RMSEA = .057), with some indices (e.g., RMSEA) suggesting a good fit to the data and others (e.g., CFI) indicative of a rather poor fit. This was somewhat surprising given the very clean factor structure suggested by the EFA results. Nevertheless, all items loaded significantly on their intended factor and the modification indices did not reveal any consistent sources of misfit. The low values of some of the fit indicators may be due to the fact that, like the WPS, the IFQ scales are multifaceted. The large number of single-item indicators per latent factor (i.e., 13 to 21), and the modest ratio of sample size to indicators (i.e., about 7:1) also might have contributed to the mixed support for the 6-factor model.

Subgroup differences. Next, we examined subgroup differences for the IFQ scale scores. In general, results were similar to those found for the WPS scales. Specifically, male recruits had significantly higher Realistic interests, whereas female recruits had significantly higher Artistic, Social, and Conventional interests (see Table 13.6). As for race/ethnicity, White recruits had significantly higher Realistic interests than both Black and Hispanic recruits, whereas Black recruits had higher Social, Enterprising, and Conventional interests than White recruits (see Table 13.7). Hispanics also had higher Conventional interests than Whites.

Fit indices. As with the WPS, we calculated D^2 and r fit indices to assess the correspondence between recruits' IFQ profiles and the environment-side profiles based on the AES and FAES data. Because IFQ items are rated on a 3-point scale and AES and FAES ratings were made on a 5-point scale, the environment-side measures were recoded to a 3-point scale such that 1 = 1, 2 = 1, 3 = 2, 4 = 3, and 5 = 3.

Descriptive statistics and intercorrelations for the IFQ fit indices are shown in Table 13.8. Recoding the AES and FAES scores did not appear to reduce relations between current and future Army fit indices (e.g., $r = .97$ for the D^2 indices). As with the WPS, the two types of fit indices were only moderately related. Also consistent with the WPS results was that there was wide variation in fit across recruits (e.g., IFQ-AES $r = -.95$ to .96). At the same time, the average magnitude of IFQ fit indices was notably smaller than the corresponding fit indices for the WPS. For instance, the mean WPS-FAES r was .31, whereas the mean IFQ-FAES r was only .06. Also note the modest level of association between IFQ and WPS fit indices (e.g., $r = .35$ for current Army D^2 indices).

We also examined subgroup differences for the IFQ fit indices (see Tables 13.9 and 13.10). Results of these analyses were highly similar to the WPS subgroup results. Specifically, the few statistically significant mean differences we found favored the majority groups (i.e., male and White recruits), but the effect sizes for those mean differences were rather modest (i.e., all $d \leq 0.30$).

Discussion

Results of the IFQ data analyses were, in general, quite positive. For example, the IFQ scales demonstrated very good measurement characteristics, and the overlap between some of the scales (e.g., Social and Enterprising) appears to be somewhat less of an issue than for the WPS. However, as with the WPS, we found some notable subgroup differences on the IFQ scale scores (and to a lesser extent on some IFQ fit indices) that favored majority group members.

We wanted to administer both the IFQ and the WPS in the concurrent validation to compare how these two types of interest measures relate to the attitudinal precursors of attrition measured by the ALS. Administration time limitations, however, will prevent administration of both interest measures to the MOS-specific samples so those Soldiers will only get the WPS; Soldiers in the Army-wide sample will get both types of measures. Also, DMDC recently updated the instrument (now called the Career Exploration Program Interest Inventory, or "CEP") on which the IFQ was based. Because the Army is unlikely to maintain two versions of the same general instrument, we decided to include the CEP rather than the IFQ in the concurrent validation data collection. Because several of the changes made to the CEP reflect design decisions that were incorporated into the IFQ (e.g., using only activities items), we expect the CEP to perform similarly to the IFQ.

Person-Side Expectations Measure: Pre-Service Expectations Survey

Description of Measure

The final measure of vocational interests is the Pre-Service Expectations Survey (PSES). The PSES is identical in format to the AES, FAES, and JCS, but includes different instructions. Specifically, respondents are asked to rate the extent to which they think the Army will provide work activities and training opportunities associated with each RIASEC dimension.[64] These data will be used to determine the extent to which the expectations of new recruits are consistent with what the current Army actually supports (assessed via the AES) for assessing expectations-reality fit. As discussed, we also believe expectations about the type of work interests the Army work environment supports might interact with measures of P-E fit to predict criteria such as attrition.

The 30-item PSES was administered to new recruits during pilot testing and in the faking research (honest condition only). No substantive changes were made to the instrument based on data analysis and qualitative feedback from respondents.

[64] The PSES also includes six more items (one per dimension) than the environment-side interest measures (i.e., 30 versus 24 total items). These items ask respondents to rate the extent to which first-term Soldiers have opportunities to develop skills associated with each RIASEC dimension.

Field Test Results

Sample

A total of 363 new recruits were administered the PSES during the predictor field test. Unlike the WPS and IFQ, recruits completed the PSES on an "as time permitted basis." As a result, only about half of the field test participants were given this instrument. Of these, 26 recruits failed to complete at least 90% of the items, and 11 respondents had little or no variation in their responses. Examination of the predictor field test problem logs did not suggest any problems. Thus, the PSES analyses use data from 324 recruits, or 89.3% of the initial sample.

Scale Refinement

The same general types of analyses used to evaluate the psychometric quality of previously discussed interest measures were used to evaluate the PSES. The analyses did not reveal any problematic items, and the items that comprise each PSES scale were best described by a single factor. Further, EFA of all 30 PSES items provided strong evidence for the intended 6-factor model.[65]

Descriptive Statistics and Reliability Estimates

Table 13.11 presents descriptive statistics and reliability estimates for the PSES scale scores. Note that the PSES scale means were much more differentiated than the means for the interest needs measures (i.e., the WPS and IFQ). Reliability estimates for the PSES scales were acceptable (i.e., .75 and higher). Correlations among the scale scores were mostly modest, ranging from .01 between Artistic and Conventional to .52 between Investigative and Artistic.

Table 13.11. Descriptive Statistics, Reliability Estimates, and Intercorrelations for PSES Scale Scores

Scale	*M*	*SD*	1	2	3	4	5	6
1 Realistic	3.95	0.61	(.75)					
2 Investigative	3.37	0.74	**.38**	(.81)				
3 Artistic	2.82	0.94	**.13**	**.52**	(.89)			
4 Social	3.64	0.74	**.35**	**.34**	**.32**	(.83)		
5 Enterprising	3.66	0.71	**.36**	**.22**	.12	**.41**	(.80)	
6 Conventional	3.93	0.75	**.41**	.11	.01	**.51**	**.51**	(.86)

Note. $n = 324$. All scales comprise five items, each of which was rated on a 5-point Likert-type scale. Internal consistency reliability estimates (alpha) are shown along the diagonal in parentheses. Bolded correlations are statistically significant, $p < .05$ (two-tailed).

Relations between Interests and Expectations

Next, we investigated relations between recruits' vocational interests and their expectations about the extent to which the Army work environment will support those interests. Table 13.12 displays zero-order correlations between PSES scale scores and the corresponding

[65] As with the AES, this finding should be interpreted with caution given that the items comprising each PSES scale were not independent (i.e., items were linked to a description of a specific RIASEC dimension).

WPS and IFQ scores. Interestingly, there was little or no relationship between recruits' needs and expectations (median $r = .09$). This was somewhat surprising given that individuals are thought to seek organizations they believe will support their interests (Schneider, 1987). Given the satisfactory psychometric characteristics of the interest measures, it seems unlikely that the small correlations are due to measurement error. Moreover, the moderately strong relations between PSES and AES scale scores (discussed later) suggest that, on average, recruits' expectations about the Army work environment are quite accurate. One possible explanation is that unlike many civilian occupations, individuals often join the Army as a means to an end (e.g., to fund their education) rather than to make the Army a career.[66] Therefore, many individuals may join the Army (e.g., for only one or two terms of service) knowing that the work environment might not support all of their vocational interests.

Table 13.12. Correlations between PSES Scale Scores and Corresponding WPS and IFQ Scores

Scale	WPS r	IFQ r
Realistic	**.13**	.06
Investigative	.05	.08
Artistic	.07	.03
Social	**.15**	**.16**
Enterprising	.10	-.04
Conventional	.09	.09

Note. $n = 303$. Statistically significant correlations sizes are bolded, $p < .05$ (two-tailed).

Subgroup Differences

As with the needs measures, we examined subgroup differences on the PSES scale scores with regard to gender and race/ethnicity. Results of these analyses are displayed in Tables 13.13 and 13.14. Only one statistically significant difference emerged for gender, namely that female recruits expected the Army to support Conventional interests to a greater extent than did male recruits. The effect size for this differences, however, was rather modest ($d = 0.36$). None of the subgroup comparisons were statistically significant for race, and like gender, what little mean differences there were tended to favor minority recruits.

Table 13.13. PSES Scale Scores by Gender

		Male		Female	
Scale	d_{FM}	M	SD	M	SD
Realistic	0.00	3.95	0.62	3.95	0.59
Investigative	0.01	3.37	0.76	3.37	0.70
Artistic	0.08	2.79	0.92	2.87	0.98
Social	0.19	3.59	0.71	3.73	0.78
Enterprising	0.14	3.63	0.74	3.73	0.65
Conventional	**0.36**	3.84	0.77	4.12	0.67

Note. $n_{Male} = 215$, $n_{Female} = 108$. Scale scores are based on items rated on a 5-point Likert-type scale. d_{FM} = Effect size for Female-Male mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., Males) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

[66] Career intentions data collected during the criterion field test (via the ALS) provide support for this assertion.

Table 13.14. PSES Scale Scores by Race/Ethnic Group

			White		Black		White Non-Hispanic		Hispanic	
Scale	d_{BW}	d_{HW}	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Realistic	0.07	-0.16	3.93	0.64	3.97	0.56	3.94	0.64	3.84	0.62
Investigative	0.12	-0.15	3.36	0.76	3.46	0.73	3.38	0.75	3.27	0.82
Artistic	0.21	0.24	2.76	0.94	2.96	1.03	2.74	0.93	2.96	1.00
Social	0.07	0.11	3.61	0.76	3.66	0.75	3.61	0.75	3.69	0.86
Enterprising	-0.19	-0.05	3.68	0.72	3.54	0.73	3.69	0.71	3.65	0.80
Conventional	-0.03	-0.17	3.93	0.78	3.90	0.75	3.95	0.77	3.82	0.84

Note. $n_{\text{White}} = 216$. $n_{\text{Black}} = 49$. $n_{\text{White Non-Hispanic}} = 193$. $n_{\text{Hispanic}} = 34$. Scale scores are based on items rated on a 5-point Likert-type scale. d_{BW} = Effect size for Black-White mean difference. d_{HW} = Effect size for Hispanic-White Non-Hispanic mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., White) are listed second in the effect size subscript. All effect sizes are nonsignificant, $p > .05$ (two-tailed).

Fit Indices

Finally, we computed fit indices to determine the extent to which recruits' expectations regarding the vocational interests the Army work environment supports (measured by the PSES) were consistent with the interests the Army actually supports (measured by the AES).[67] Descriptive statistics and intercorrelations for the PSES fit indices can be found in Table 13.15. It is noteworthy that the overall level of fit was higher for the PSES than for interest needs measures (see Table 13.8). For example, the mean PSES-AES *r* (.51) was much higher than the corresponding *r*s for the WPS (.31) and IFQ ($r = -.01$). Nonetheless, there was still wide variation in fit indices across recruits, which suggests that some recruits have very realistic expectations about the interests supported by the current Army work environment while other recruits do not.

Table 13.15. Descriptive Statistics and Intercorrelations for PSES Fit Index Scores

Fit Index	*M*	*SD*	1	2
1 Current Army D^2	3.59	3.14	–	
2 Current Army *r*	0.51	0.43	-.39	–

Note. $n = 658$. Fit indices are based AES profile scores. All correlations are statistically significant, $p < .05$ (two-tailed).

Tables 13.16 and 13.17 present subgroup descriptive statistics for the PSES fit indices. As shown, there were no statistically significant subgroup mean differences with regard to gender or race/ethnicity.

[67] We did not compute fit indices between the person-side expectations measures and the future Army environment-side measures because we do not believe relations between recruits' expectations about the current Army and SME ratings about of the future Army to be conceptually meaningful. Specifically, with the PSES, recruits indicate their expectations only about the current Army environment, not what they expect the future environment will be like.

Table 13.16. PSES Fit Index Scores by Gender

		Male		Female	
Fit Index	d_{FM}	*M*	*SD*	*M*	*SD*
Current Army D^2	-0.02	3.62	3.29	3.54	2.85
Current Army *r*	0.13	0.50	0.45	0.55	0.38

Note. n_{Male} = 215, n_{Female} = 108. Fit indices are based AES profile scores. d_{FM} = Effect size for Female-Male mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., Males) are listed second in the effect size subscript. All effect sizes are statistically nonsignificant, $p > .05$ (two-tailed).

Table 13.17. PSES Fit Index Scores by Race/Ethnic Group

			White		Black		White Non-Hispanic		Hispanic	
Fit Index	d_{BW}	d_{HW}	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Current Army D^2	0.03	0.19	3.73	3.39	3.82	2.91	3.66	3.14	4.24	4.50
Current Army *r*	-0.22	-0.29	0.53	0.43	0.43	0.42	0.54	0.42	0.42	0.52

Note. n_{White} = 216. n_{Black} = 49. $n_{White\ Non\text{-}Hispanic}$ = 193. $n_{Hispanic}$ = 34. Fit indices are based AES profile scores. d_{BW} = Effect size for Black-White mean difference. d_{HW} = Effect size for Hispanic-White Non-Hispanic mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., White) are listed second in the effect size subscript. All effect sizes are statistically nonsignificant, $p > .05$ (two-tailed).

Discussion

The PSES appears to be a promising measure for assessing expectations-reality fit. For example, its scales exhibited sufficient variation and acceptable levels of internal consistency reliability. The overall lack of significant subgroup differences (for both scale scores and fit indices) is another attractive feature of this measure.

As with the other expectations measures discussed in this chapter, the PSES will not be administered in the concurrent validation because the sample will comprise Soldiers 18 to 36 months of service. As such, it would not be appropriate to ask these Soldiers about what vocational interests they "expect" the Army work environment to support. Indeed, the PSES was designed as a pre-enlistment assessment to identify potential recruits with inaccurate expectations about the Army work environment. Data from this measure will be included in the Select21 attrition database. As this database matures, we will examine relations between attrition and PSES scale scores and fit indices.

WORK VALUES MEASURES

We now shift our discussion to measures that assess work values. As with the vocational interest measures, we begin with the environment-side supplies measure.

Environment-Side Supplies Measure: Army Description Inventory

Description of Measure

The Army Description Inventory (ADI) is a 42-item instrument that assesses the extent to which the Army supports Soldiers' work values during their first-term of service. The Army supports work values by providing occupational or organizational reinforcers, which are defined as the environmental stimulus conditions (e.g., the Army's provision of opportunities to learn new skills) associated with persons' work values (Dawis & Lofquist, 1984). Persons' work values, in turn, are revealed through the importance they place on various reinforcers that may or may not be supplied by a work environment. The reinforcers assessed by the ADI were identified through a review of the work values literature. Twenty-one of the reinforcers included in the ADI were derived from the Dawis and Lofquist (1984) taxonomy of occupational reinforcers. The other 21 reinforcers were created specifically for Select21. Of these additional reinforcers, 17 represented completely new content, and four were re-wordings of reinforcers from Dawis and Lofquist. The 17 new reinforcers resulted from a review of (a) the general literature on work values (e.g., Schwartz, 1994), (b) recent research on the values of American youth (Sackett & Mavor, 2002), (c) ARI's Army Values study (Ramsberger, Wetzel, Sipes, & Tiggle, 1999), and (d) the Select21 job analysis results. These new reinforcers were added to help round out the Dawis and Lofquist taxonomy for use in the Army context.

The ADI was designed to be administered to Army SMEs. It presents SMEs with a description of each reinforcer (e.g., first-term Soldiers in the Army learn new skills) and asks them to indicate their level of agreement with regard to whether the Army provides the reinforcer to first-term Soldiers. SMEs are asked to make their ratings on a 5-point Likert-type scale with anchors that range from *strongly disagree* (1) to *strongly agree* (5). The purpose of the ADI is to create an Army reinforcer profile against which applicant work value profiles can be compared to assess P-E fit. Specifically, we will compare mean SME ratings on each ADI reinforcer to recruits' ratings on the Work Values Inventory (WVI) to assess needs-supplies fit, and to recruits' ratings on the Army Beliefs Survey (ABS) to assess expectations-reality fit.

Results

The ADI was administered to two groups of SMEs: (a) 69 NCO Drill Sergeants and AIT instructors (E5-E7), and (b) six members of the Select21 SME Panel during the pilot tests. The first group was asked to complete the ADI as it pertains to the current Army, whereas the latter group was asked to complete the ADI with regard to what the Army will offer future first-term Soldiers in light of anticipated future conditions.

Prior to conducting analyses with the "current Army" ADI data, we screened the data for individual SMEs who had ratings that were highly inconsistent with ratings of the "average"

SME.[68] To check the data for outlier SMEs, we first standardized ratings for each of the 28 reinforcers in the field test version of the Work Values Inventory (WVI).[69] Next, we calculated the average absolute *z*-score for each SME across the 28 reinforcers. Lastly, we removed any SME whose average absolute *z*-score was more than 1.64 *SD*s above the average absolute *z*-score across SMEs. Based on this criterion, data for six SMEs were removed from the sample.

Table 13.18 displays descriptive statistics for ratings on the 28 ADI reinforcers retained for the field test WVI. ICCs were used to estimate the consistency with which SMEs rank-ordered the reinforcers. The resulting single- and *k*-rater reliability estimates, respectively, were .40 and .98 for current Army ratings and .48 and .84 for future Army ratings.

As for interrater agreement with regard to the provision of individual reinforcers, *SD*s were generally small, with raters tending to agree more on reinforcers with high mean ratings. The fact that the *SD*s were smaller for highly related reinforcers is partially a function of range restriction (recall the upper bound on the ADI rating scale is 5). As discussed earlier in the chapter, the *SD*s for uniformly and normally distributed ratings on a 5-point scale are 1.41 and 1.15, respectively. Because most of the observed *SD*s are notably smaller than these values, it suggests that raters tended to agree with regard to the Army' provision of these reinforcers.

The ordering of reinforcers within current and future Army profiles was consistent with our expectations. For example, SMEs indicated that the current and future Army environments will tend to offer Soldiers opportunities to establish friendships with co-workers (Co-Workers), advancement (Advancement), work as a team (Team Orientation), learn new skills (Skill Development), help others (Social Service), gain personal discipline and maturity (Emotional Development), and improve their physical fitness (Physical Development). Conversely, SMEs indicated that the current and future Army environments will not tend to support Soldiers' needs for planning their work with little supervision (Autonomy), working alone (Independence), settling down in one location for an extended period (Home), or having a flexible work schedule (Flexible Schedule). Given these similarities, it was not surprising that the correlation between current and future Army profiles was quite high ($r = .88$). These findings suggest that the future Army is generally going resemble the current Army in terms of the opportunities it does and does not offer first-term Soldiers.

Although the current and future Army ADI profiles were quite similar, there were differences between the current and future Army ratings on individual reinforcers. For example, ratings of reinforcers categorized as "high supply" in the current Army were generally seen as being in even greater supply in the future Army (mean $d = -0.61$).[70] In contrast, small differences were found between current and future ratings for reinforcers categorized as "low supply" (mean $d = 0.13$). As discussed in the vocational interests section, large differences between current and

[68] Given the small number of SMEs who provided "future Army" ADI ratings, their data was not subject to the same screening process. Based on visual inspection of the future-oriented ADI data, no SMEs were eliminated.
[69] Although the ADI included 42 reinforcers, we subsequently assessed only a subset of these in the person-side work values measures (i.e., the WVI and ABS). We discuss the process used to identify this subset of reinforcers later in the chapter.
[70] As discussed later, for constructing pre-field test versions of the WVI, ADI reinforcers were classified into three categories that corresponded to the degree to which the Army "supplies" them to first-term Soldiers.

future Army ratings on individual dimensions has implications for how we plan to combine person and environment data for prediction of the Select21 criteria (see Appendix I).

Table 13.18. Descriptive Statistics for ADI Scale Scores

		Current Army					Future Army					
Scale	Supply Category	Rank	*M*	*SD*	SE_M	p_{NCO} >= 4	Rank[a]	*M*	*SD*	SE_M	p_{NCO} >= 4	*d*
Co-Workers	High	1	4.33	0.52	0.08	0.98	1	4.67	0.52	0.21	1.00	-0.65
Advancement	High	2	4.30	0.52	0.07	0.97	2	4.50	0.55	0.22	1.00	-0.38
Feedback	High	3	4.09	0.41	0.06	0.96	9	4.33	0.52	0.21	1.00	-0.59
Emotional Development	High	4	4.06	0.63	0.09	0.84	3	4.50	0.55	0.22	1.00	-0.70
Achievement	High	5	4.05	0.68	0.09	0.89	15	4.00	0.63	0.26	0.83	0.07
Social Service	High	6	4.00	0.56	0.08	0.89	10	4.33	0.52	0.21	1.00	-0.59
Physical Development	High	7	3.96	0.76	0.11	0.82	5	4.50	0.55	0.22	1.00	-0.71
Team Orientation	High	8	3.93	0.65	0.10	0.89	4	4.50	0.55	0.22	1.00	-0.88
Skill Development	High	9	3.92	0.53	0.08	0.82	6	4.50	0.55	0.22	1.00	-1.09
Fixed Role	Mid	10	3.84	0.77	0.11	0.78	18	3.50	0.84	0.34	0.67	0.44
Travel	Mid	11	3.84	1.03	0.15	0.67	12	4.17	0.98	0.40	0.67	-0.32
Recognition	Mid	12	3.78	0.74	0.09	0.75	11	4.33	0.52	0.21	1.00	-0.74
Social Status	Mid	13	3.72	0.77	0.12	0.67	8	4.50	0.55	0.22	1.00	-1.01
Societal Contribution	Mid	14	3.72	0.83	0.12	0.65	7	4.50	0.55	0.22	1.00	-0.94
Leisure Time	Mid	15	3.65	0.69	0.10	0.76	19	3.00	1.41	0.58	0.33	0.94
Leadership Opps.	Mid	16	3.52	0.81	0.12	0.67	20	3.00	1.41	0.58	0.33	0.64
Supportive Supervision	Mid	17	3.51	0.80	0.10	0.59	16	4.00	0.63	0.26	0.83	-0.61
Ability Utilization	Mid	18	3.50	0.84	0.10	0.63	13	4.17	0.41	0.17	1.00	-0.80
Activity	Mid	19	3.20	1.04	0.13	0.39	17	3.83	1.17	0.48	0.67	-0.61
Esteem	Mid	20	3.20	0.88	0.13	0.43	14	4.17	0.41	0.17	1.00	-1.10
Creativity	Low	21	3.06	0.88	0.11	0.33	21	3.00	1.26	0.52	0.50	0.07
Variety	Low	22	2.98	0.90	0.11	0.33	22	3.00	0.89	0.37	0.33	-0.02
Influence	Low	23	2.72	0.91	0.13	0.22	24	2.50	1.22	0.50	0.33	0.24
Comfort	Low	24	2.72	0.81	0.12	0.15	25	2.50	1.05	0.43	0.17	0.27
Flexible Schedule	Low	25	2.65	0.99	0.14	0.24	26	2.17	0.98	0.40	0.17	0.48
Autonomy	Low	26	2.41	0.71	0.09	0.08	27	2.17	0.98	0.40	0.17	0.34
Independence	Low	27	2.38	0.85	0.11	0.16	23	2.67	0.82	0.33	0.17	-0.34
Home	Low	28	2.13	1.00	0.14	0.10	28	2.17	1.17	0.48	0.17	-0.04

Note. n = 43-63 for current Army ratings and n = 6 for future Army ratings. Supply category = category that value was placed in for pre-field test versions of the WVI (based on current Army ratings). Rank = rank of mean rating. $p <= 4$ = proportion of raters who agreed or strongly agreed that the Army environment provided the given reinforcer. d = Standardized mean difference (calculated by subtracting the mean future Army ratings from the mean current Army ratings and dividing by the standard deviation of the current Army ratings).
[a]In the case of ties among mean future Army ratings, higher rankings were given to dimensions with higher mean current Army ratings.

Next Steps for the ADI

As with the vocational interest environment-side measures (i.e., AES and FAES), we will not collect further data on the ADI. The purpose of the ADI was to create an Army reinforcer profile against which recruits' ratings on the WVI and ABS could be compared to assess needs-supplies, and expectations-reality fit with regard to work values, respectively. Sufficient data to

create the ADI profiles for subsequent efforts have been obtained. Should the Army decide to adopt either the WVI or ABS for operational use, the ADI should be periodically re-administered to SMEs to assess potential changes in the Army work environment in terms of its provision of reinforcers.

Person-Side Needs Measure: Work Values Inventory

Description of Measure

The WVI is designed to assess recruits' work values. Work values are indicated by the importance individuals place on various reinforcers (e.g., the opportunity to learn new skills). The WVI is a computerized assessment consisting of four parts. The assessment asks recruits to rank order 28 reinforcers in terms how important they would be in their ideal job, and distinguish between important and unimportant reinforcers (in an absolute sense). In the first part of the WVI, respondents are asked to sort 28 reinforcers into four categories of varying importance. For example, respondents place their seven most important reinforcers in Category A and their seven least important reinforcers in Category D. Respondents then rank-order the importance of the reinforcers within each category. After completing their rankings within each category, respondents are presented with the full list of reinforcers in the order they ranked them. Upon reviewing this list, they are asked to make a line through it—above the line are reinforcers they deem important to have on their ideal job, and below the line are reinforcers they deem unimportant to have on their ideal job.

Prior to the field test, the WVI was formatted quite differently than described above. Changes were made to the WVI based on results from pilot test and faking research data collections. In the sections that follow, we describe (a) development of the original WVI, (b) its scoring, and (c) changes made for the field test.

Pre-Field Test WVI

Description of Measure

Given that virtually all reinforcers included on the ADI are fairly desirable (e.g., having friendly co-workers), asking respondents to use a simple Likert-type to assess the degree to which they value each reinforcer would not likely yield useful information. As such, we adopted a forced-choice format for the original WVI.

Based on ADI data, we identified 27 work values to assess in the pre-field test WVI.[71] The choice to assess only 27 values was due to concerns about testing time and redundant content among the reinforcers. An important characteristic of the 27 values assessed by the pre-field test WVI was that the reinforcers they corresponded to could be classified into one of three "supply" categories: (a) a "high" category reflecting reinforcers supplied by the Army to first-term Soldiers, (b) a "low" category reflecting reinforcers *not* generally supplied by the Army,

[71] The pre-field test WVI did not assess four values that were eventually included in the field test WVI, and assessed three values that were excluded from the field test WVI. Reasons for these changes are discussed later in this chapter.

and (c) a "mid" category reflecting reinforcers that fall somewhere between reinforcers in the high and low categories in terms of the degree to which the Army supplies them. As discussed later, grouping reinforcers into these categories was essential for creating a forced-choice measure that covered as many of the reinforcers assessed by the ADI, yet did so with a reasonable number of items.

We used a two-step process to reduce the 42 ADI reinforcers to 27 and sort them into the aforementioned supply categories.[72] First, we formed initial sets of high, low, and mid supply reinforcers based on mean ADI ratings, as well as criterion-related validity evidence from the Project A work values measure, the Job Orientation Blank (J. P. Campbell & Knapp, 2001). After this step, there were 18, 12, and 12 reinforcers, respectively in the high, low, and mid categories. Next, we reduced the number of reinforcers in each of these sets to nine. Reducing these initial sets was based on empirical considerations (e.g., mean ADI ratings) and rational judgments by the authors. The final set of reinforcers in each category reflected those we thought would perform best in an operational context (e.g., retain their criterion-related validity) and remain stable in terms of their supply by the Army during the transition to the Future Force. Of the 27 selected reinforcers, 14 were based on the Dawis and Lofquist (1984) taxonomy and 13 were new reinforcers created for Select21.

Next, we constructed the pre-field test version of the WVI, which included 81 forced-choice items that assessed the 27 work values. Each item comprised three reinforcers (i.e., a triad). Respondents were asked to indicate which reinforcer in each triad was most and least important to them. The triads were structured so that each contained (a) one "high supply" reinforcer (k = 9), (b) one "low supply" reinforcer (k = 9), and (c) one "mid supply" reinforcer (k = 9). The WVI was constructed such that no two reinforcers were paired together more than once across the entire instrument, and reinforcers from the same category were never paired together. For example, reinforcers in the low supply category were only paired with reinforcers in the high and mid supply categories, never with each other. This format meant that each reinforcer was compared to 18 other reinforcers. This selective grouping of reinforcers was done out of necessity, as 2,925 triads would be needed to compare all possible combinations of 27 reinforcers. Nevertheless, this format allowed us to assess the extent to which respondents prefer reinforcers the Army supplies to those it does not supply, and vice versa. Figure 13.3 shows a sample item from the pre-field test WVI.

Indicate which statement is most important to you in your ideal job (the kind of job you would most like to have), and which statement is least important to you in your ideal job.

On my ideal job, I would...
- (a) receive recognition for what I do.
- (b) gain personal discipline and maturity.
- (c) have a flexible work schedule.

Figure 13.3. Sample item from the pre-field test WVI.

[72] See Knapp (2003) for further details.

Scoring the Measure

To score the pre-field test version of the WVI, we calculated separate work value scale scores corresponding to each of the 27 reinforcers. The algorithm used to generate these scale scores involved calculating the proportion of times each reinforcer was preferred over reinforcers in the other supply categories. Based on simple variations on this algorithm, three types of scores were generated for each reinforcer. These scores differed according to what reinforcers served as *referents* (i.e., the reinforcers against which the reinforcer of interest was compared). For example, we generated three value scale scores for reinforcers in the high supply category by calculating the proportion times they were (a) preferred over reinforcers in the mid supply category, (b) preferred over reinforcers in the low supply category, and (c) preferred over all non-high supply reinforcers. Conversely, we generated three scale scores for reinforcers in the low supply category by calculating the proportion of times they were (a) preferred over reinforcers in the mid supply category, (b) preferred over reinforcers in the high supply category, and (c) preferred over all non-low supply reinforcers. The rationale for this referent-based algorithm was based on two hypotheses. First, recruits who prefer reinforcers that are in high supply in the Army over reinforcers that are in lower supply in the Army will have more positive attitudes towards the Army. Second, recruits who prefer reinforcers that are in low supply in the Army over reinforcers that are in greater supply in the Army will have more negative attitudes towards the Army.

Pilot Test Results

The pre-field test WVI was administered to over 400 new recruits during pilot testing at the reception battalions. These pilot test data provided a basis for making slight wording changes to the measure, and getting a general idea for how the measure functioned (e.g., clarity of instructions to recruits). In general, the WVI value scales appeared to exhibit a good amount of variation, and only minor changes were made to the measure in preparation for the faking research.

Faking Research Results

After pilot testing the WVI, we administered it to new recruits in the faking research data collections. The WVI was administered under honest and "faking" conditions to assess its susceptibility to response distortion. A mixed between-within subjects design was used whereby recruits first completed the WVI (n = 193) under honest instructions. Then, about half of the recruits (n = 93) were instructed to fake as much as they could (i.e., the "fake max" condition) and others (n = 100) were instructed how to fake (i.e., the "coached" condition). For the latter condition, recruits were instructed to indicate the reinforcer in each triad that sounded "most like the Army" was most important to them and indicate that the reinforcer in each triad that sounded "least like the Army" was least important to them. In sections below, we briefly describe results from the faking research and how they affected the decision to revise the WVI.

A key question we wanted to address with the faking data regarded the extent to which recruits were able to identify the high and low supply reinforcers in each triad (i.e., fake effectively). Such a pattern of responding is indicated by an inflation of scores for high supply reinforcers and deflation of scores for low supply reinforcers under faking conditions. This response set could potentially lead to "invalid" scores on the WVI if it does not reflect recruits'

true preferences. We were aware of this possibility when designing the WVI, but hypothesized that other design characteristics of the WVI would mitigate the impact of such a response pattern on criterion-related validity. Specifically, we hypothesized that for respondents to inflate their scores on high supply reinforcers and deflate their scores on low supply reinforcers, their expectations of what the Army supplied (and did not supply) would need to be accurate. Given the relationship between accuracy of expectations and job attitudes (Wanous, 1992), we hypothesized that this contaminate source of variation in the assessment of work values would not detract from criterion-related validity. In other words, recruits' expectations might introduce contaminate variation into the WVI (thus detracting from its construct-related validity), but it would be variation that covaries with criteria of interest.

Analyses of the faking research data indicated that instructing recruits to fake notably inflated the proportion of times they preferred high supply reinforcers over low supply reinforcers (Honest: 67%, Fake Max: 77%, Coached: 82%), and conversely, deflated the proportion of times they preferred low supply reinforcers over high supply reinforcers (Honest: 34%, Fake Max: 23%, Coached: 18%).[73] Interestingly, although there were elevation differences between honest and faking conditions, there were only minimal differences in *SD*s across conditions. Thus, the WVI was still able to differentiate among recruits in the faking conditions.

These findings led us to pose another question. Specifically, if variation was maintained across conditions, was the nature of that variation the same? If so, then recruits would be rank-ordered similarly across conditions. In a directed faking study, motivations to fake are likely equalized far more than they are in practice, and thus should have little impact on the rank ordering of respondents across conditions. Thus, any lack of relationship between honest and faked scores such should reflect (a) differences in ability to fake effectively, and/or (b) differences in compliance with the instruction sets. To investigate these possibilities, we examined the correlation between respondents' honest and faked scores. As with the vocational interests measures (i.e., the WPS and IFQ), correlations between honest and faked responses were generally quite low (e.g., $r = .01$ to $.25$ for the low and high supply composites). These results suggest that there are notable individual differences in recruits' ability to fake the WVI effectively and/or their compliance with the faking instructions.

In developing the WVI, we hypothesized that recruits who were "able to fake effectively" would have realistic expectations regarding the Army work environment (thus, potentially contributing criterion-relevant contamination to the WVI). To investigate this possibility, we correlated recruits' honest and faked WVI scale scores with corresponding scale scores on the Army Beliefs Survey (ABS). The ABS (discussed later in this chapter) was designed to assess recruits' expectations regarding the extent to which the Army supplies the WVI reinforcers. Results of these analyses indicated that expectations were more related to WVI scores obtained under the coaching (mean $r = .21$ for low supply reinforcers and .23 for high supply reinforcers) and fake maximum conditions (mean $r = .16$ for low supply reinforcers and .17 for high supply reinforcers) than to WVI scores obtained under honest conditions (mean $r = .01$ for low supply reinforcers and .12 for high supply reinforcers). These findings partially account for the lack of correlation observed between honest and faked scores. Specifically, a statistically significant portion of variation in faked scores was accounted for by factors that reflect individual

[73] Appendix J provides detailed faking research results at the level of individual reinforcers.

differences in recruits' expectations, whereas expectations appeared relatively unrelated to honest scores (particularly among the low supply reinforcers).

WVI for the Predictor Field Test

Modifications to WVI Format

The decision to change the format of the WVI for the field test was based on several factors. First, the administration time of the pre-field test version of the WVI was far too long (e.g., it took 35-40 minutes to complete during the faking research). For the field test, we were limited to a 30-minute administration time. As such, we had to either change formats or reduce the number of values examined. For reasons noted below, we opted to change formats.

A second reason for changing formats was that we believe pairing reinforcers from different supply categories in each triad exacerbated the transparency of the response set discussed earlier (i.e., endorsing high supply reinforcers as most important, and low supply reinforcers as least important). Moving to a rank-order format would likely make it more difficult for respondents to engage in this response set. Specifically, by not using the triads, we are making the contrast between high and low supply reinforcers less salient, thus potentially attenuating respondents' ability to respond in the manner described above.

A third reason for changing formats was based on relations between Basic Combat Training (BCT) attrition and expectations data observed during the pilot test and faking research. Specifically, we found that expectations were not consistently related to BCT attrition. Given that expectations appear to contaminate faked WVI scores, to the extent that such contamination occurs in practice it may attenuate relations between the WVI and early attrition.[74] We felt that changing formats would make responding to the WVI using an expectations-based response set more difficult for applicants. Although these findings contributed to our decision to change formats, it was only one of many contributing factors. We emphasize this because the attrition data we used were limited to BCT attrition. We hypothesize that expectations will be more related to later attrition (e.g., unit attrition), after Soldiers have been in the Army long enough for their attitudes to develop.

A final reason for changing the format of the WVI involved limitations that the pre-field test format put on our ability to capitalize on the ADI data. Changing formats not only allows us to continue to evaluate the referent-based algorithm used for the pre-field test WVI, but also allows for a new scoring algorithm that is independent of the environment-side data (discussed later).[75] A beneficial feature of the new algorithm is that it gives us greater flexibility with regard to how we combine WVI and ADI data to predict relevant criteria. Specifically, it will allow us to assess different ways of combining person and environment information for predicting criteria.

---

[74] The elevation of scores in practice could be less than that observed in the faking research. Recall that respondents in the faking sample were new recruits about to enter training. As such, they may have more accurate expectations of what the Army is like, which, in turn may have enhanced their ability to fake the WVI relative to applicants.
[75] Under the referent-based algorithm, the way in which person- and environment-side data are combined is pre-determined (i.e., it is reflected in the choice of what reinforcers are compared to one another).

Modifications to WVI Content

We also made changes to the WVI content based on analysis of pre-field test data and results of ongoing ARI attrition research (e.g., Putka & Strickland, 2004; Strickland, 2004). For the field test, we dropped three reinforcers from the pre-field test WVI (Security, Compensation, and Company Policies and Procedures). The decision to drop these reinforcers was based on one or more of the following criteria: (a) highly skewed response distributions, (b) large honest-fake effect sizes, (c) low correlations with early attrition data, (d) low correlations with attitudinal data (e.g., satisfaction) collected from first-term Army Soldiers early in the project, (e) highly skewed ABS response distributions, (f) low ABS validity for predicting early attrition, and (g) research suggesting that such values might not predict later attrition (e.g., Strickland, 2004). Although these reinforcers were dropped, we will still evaluate their potential utility for predicting later attrition. Specifically, as more mature attrition data become available for pilot test and faking research participants, we will re-examine the relationship between the value scales corresponding to these reinforcers and attrition.[76] If the aforementioned scales predict later attrition, we may want to consider revising the WVI content for the concurrent validation.

Finally, we also made slight wording changes to three reinforcers (Leisure Time, Fixed Role, and Influence) based on feedback from respondents, project staff, and data analysis. In addition, we added four reinforcers assessed in the ADI but dropped from the pre-field test WVI due to administration time constraints. Two of these reinforcers (Activity and Supportive Supervision) were derived from the Dawis and Lofquist (1984) taxonomy, and three (Supportive Supervision, Travel, and Physical Development) were linked to attrition in the Strickland (2004) research.

Scoring the WVI

The revised WVI offers several potential ways to generate scale-level scores. One option would be to adopt the referent-based algorithm used to score the pre-field test measure. A drawback of this algorithm is that ADI data were used to determine which reinforcers to compare to one another (e.g., high supply vs. low supply). Thus, this algorithm produces person-side value scale scores that are partially dependent on the environment-side information. This imposes constraints on how we assess P-E fit, and limits options for combining person- and environment-side data to predict relevant criteria. For example, with the referent-based scoring algorithm, it would not be meaningful to examine the similarity of WVI and ADI profile scores because the ADI data would already be reflected in the WVI scale scores. For the same reason, it would not be meaningful to examine differences between WVI and ADI scores at the level of individual reinforcers (i.e., differences between profile elements). Lastly, such a scoring algorithm would not allow us to assess whether person- and environment-side data should be combined in different ways to predict criteria as a function of the type of value and criterion examined (as discussed in Appendix I). For these reasons, we focused on an alternative algorithm for generating WVI scale scores for the field test.[77]

[76] See Putka (2004) for an overview of the attrition-related analyses planned for Select21.

[77] In this report, we present field test results for the WVI based only on the new algorithm. Nonetheless, we will still generate scores for the field test WVI using the referent-based algorithm. These scores will be included in the Select21 attrition database (Putka, 2004). This will allow us to explore the potential utility of the referent-based scoring algorithm for predicting attrition.

We plan to adopt an algorithm for scoring the WVI scales that parallels the algorithm used to score the Minnesota Importance Questionnaire (MIQ; Gay, Weiss, Hendel, Dawis, & Lofquist, 1971) and the Occupational Information Network (O*NET) Work Importance Profiler (WIP; McCloy et al., 1999). We subsequently refer to this as the MIQ/WIP algorithm.[78] The MIQ and WIP are similar to the WVI in that they involve rank ordering of reinforcers and differentiating between important and unimportant reinforcers as a final step in the assessment process. Furthermore, all of these measures draw heavily from the Dawis and Lofquist (1984) taxonomy of occupational reinforcers. This algorithm yields 28 work value scale scores for the WVI that are expressed in a *z*-score metric. Work values scale scores greater than 0 indicate a value is important to the respondent, and scale scores less than 0 indicate a value is not important to the respondent.

There are several benefits of the MIQ/WIP scoring algorithm. First, it provides a better approximation of persons' normative standing on each value than would rank-order information alone (Hicks, 1971). This is achieved by using data from the final step in the WVI assessment (i.e., differentiating between important and unimportant reinforcers) to establish an individual zero-point on each value's importance scale. Establishing such a zero-point allows for more meaningful between-person comparisons because the ipsativity of the assessment is reduced (Gay et al., 1971).

Another benefit of the MIQ/WIP scoring algorithm is that the resulting scale score for each WVI work value is independent of environment-side data. Recall that with the referent-based scoring algorithm, WVI scale scores are partially a function of environment-side ADI data. With the MIQ/WIP algorithm, we have a purer assessment of work values. Having an independent assessment of values gives us more flexibility in how we combine WVI and ADI data to predict Select21 criteria. In addition, having independent WVI and ADI scales will also allow us to report commonly used fit indices (D^2, r) for assessing profile similarity, as well as difference scores for individual work values.

One drawback of the MIQ/WIP algorithm is that we will be unable to implement it on pre-field test WVI data. Nonetheless, we will continue to evaluate both algorithms for constructing scale scores to see how well they predict the Select21 criteria.

Field Test Results

Sample

A total of 656 recruits completed the WVI as part of the predictor field test. Of these recruits, 526 (80.2%) had data that were deemed useable. Among the recruits who had useable data, 35 (6.7%) had to restart the WVI due to programming "bugs." For the majority of these recruits (51.4% of restarts), the problem arose from a bug encountered during one of the WVI's sorting tasks.[79] Fortunately, we did not lose data because of restarts. In general, recruits simply restarted the WVI and their new responses were written to the database. Prior to the concurrent validation, we will be working with programmers to fix the bugs that led to the restarts.

[78] Details of this algorithm are presented in Appendix I.
[79] Specific reasons for restarts are documented in the field test problem logs.

Unfortunately, other programming bugs led to the loss of large chunks of WVI data. By far the largest loss of data arose from problems on the final task on the WVI—differentiating between important and unimportant reinforcers using the slider bar. A set of programming bugs on this last task made it impossible to distinguish between recruits who considered either none or one of the reinforcers to be important. As a result, data for 113 recruits were deemed unusable for field test analyses (17.2% of the total WVI sample). The set of bugs on this last task accounted for 86.9% of the WVI data that were deemed unusable. Fortunately, solutions to these bugs have been identified and will be implemented for the concurrent validation.

With regard to the remaining WVI data deemed unusable, 13 recruits had missing or repeated data (e.g., two reinforcers receiving the same rank). Potential sources of the bugs that caused these errors will need to be investigated. Fortunately, these errors accounted for the loss of only 2% of the WVI data. We also eliminated data for nine recruits who completed the WVI in less than 5 minutes. Lastly, data for five recruits were eliminated due to other general reasons documented in the problems log.

Administration Time

It took recruits an average of 15.8 minutes to complete the WVI (*Mdn* = 15.1, *SD* = 5.8). A total of 90% of recruits completed the WVI in less than 24 minutes, while 95% of recruits completed it in less than 26 minutes. Recall, the administration time of the pre-field test version of the WVI was 35-40 minutes, thus the new format greatly reduced its administration time.

Descriptive Statistics

Table 13.19 shows descriptive statistics for each WVI scale scored using the MIQ/WIP algorithm. On average, recruits most preferred work that provides opportunities for Achievement, Advancement, Ability Utilization, and Comfort. Recruits expressed least preference for work that provides opportunities for Influence, Travel, and Independence. Examination of score distributions revealed that WVI scale scores were generally normally distributed.

Table 13.19. Descriptive Statistics for WVI Scale Scores

Scale	*M*	*SD*	Scale	*M*	*SD*
Advancement	1.10	0.87	Emotional Development	0.19	1.17
Achievement	0.87	0.89	Recognition	0.18	1.02
Ability Utilization	0.76	0.98	Co-Workers	0.15	0.97
Comfort	0.70	1.03	Variety	0.07	0.98
Social Status	0.61	1.14	Feedback	0.01	0.87
Leisure Time	0.50	1.05	Creativity	-0.01	0.96
Skill Development	0.48	1.01	Esteem	-0.07	0.93
Social Service	0.41	1.03	Autonomy	-0.12	0.93
Societal Contribution	0.38	1.16	Team Orientation	-0.23	1.04
Fixed Role	0.35	1.00	Activity	-0.26	1.04
Supportive Supervision	0.34	1.04	Home	-0.46	1.12
Flexible Schedule	0.24	1.04	Influence	-0.64	0.88
Physical Development	0.20	1.06	Travel	-0.80	1.18
Leadership Opportunities	0.20	1.03	Independence	-0.91	1.05

Note. *n* = 526. Scales are shown in descending order by mean score.

Subgroup differences on the WVI scales. Tables 13.20 and 13.21 show mean WVI scale scores by gender and race/ethnicity, respectively. Statistically significant gender differences were found for 9 of the 28 WVI scales. For eight of the nine differences, females exhibited greater preferences for the reinforcers than males (exception was Leisure Time). Nonetheless, the magnitudes of these effects sizes were relatively small (e.g., all but two effect sizes exceeded 0.40 in magnitude).

With regard to race/ethnicity, five statistically significant differences were found between Whites and Blacks on the WVI scales, and one statistically significant difference was found between White Non-Hispanics and Hispanics. In each case, Whites exhibited more of a preference for reinforcers than the given minority group. As was the case with gender differences, these effects sizes were relatively small, with no effect size exceeding 0.40.

Scale Intercorrelations and Factor Structure

Table 13.22 shows intercorrelations among the WVI scales. On average, the WVI scales showed modest levels of intercorrelations (mean $r = .25$). Interestingly, very few of the correlations were negative. Often when dealing with forced choice measures, many intercorrelations are negative due to the ipsativity of the data (Hicks, 1970). These results support our contention that the WIP/MIQ algorithm reduces the ipsativity of the WVI scores, and in turn, enhances the degree to which the scores provide estimates of respondents' normative standing on each WVI scale.

Next, we conducted an EFA to examine the factor structure underlying the 28 WVI scales. In the initial analysis, we did not specify the number of factors to be extracted. Results indicated that seven factors had eigenvalues over 1.0. The pattern matrix for this 7-factor solution revealed several cross loadings. As such, we conducted a series of follow-up EFAs that constrained the solution to four, five, and six factors, respectively. Comparing the results from these analyses, we adopted the 6-factor solution because of its highly interpretable factors and few cross loadings. The pattern matrix resulting from the 6-factor model, along with final eigenvalues and communalities, are shown in Table 13.23.

The factors underlying the WVI scales appear to reflect six work value constructs: (a) a need for growth (Growth), (b) a need for comfortable and flexible work environment (Comfort), (c) a need for stimulation (Stimulation), (d) a need for recognition and status (Status), (e) a need to be of service to others (Altruism), and (e) a need for self-direction (Self-Direction). Comparing this structure to the factor structure underlying MIQ and WIP reveals both similarities and differences (Dawis & Lofquist, 1984; McCloy et al., 1999). For example, the WVI Comfort, Status, Altruism, and Self-Direction factors are quite similar to the MIQ and WIP Comfort, Status, Altruism, and Autonomy factors. However, unlike the MIQ and WIP factor structures, no Achievement or Safety factors were present in the WVI data. The two factors that emerged in the WVI but not in the MIQ and WIP (Growth and Stimulation) are relevant to the Army context, as the Army tends to offer opportunities for both growth and stimulation. Also of note, the WVI factor structure appears to be cleaner (in terms of both fewer cross-loadings and interpretability) than both the MIQ and WIP factor structures presented in past research (Dawis & Lofquist, 1984, p. 83; McCloy et al., 1999, p. 50).

Table 13.20. WVI Scale Scores by Gender

Scale	d_{FM}	Male		Female		Scale	d_{FM}	Male		Female	
		M	*SD*	*M*	*SD*			*M*	*SD*	*M*	*SD*
Advancement	0.06	1.09	0.86	1.14	0.88	Emotional Development	0.15	0.13	1.19	0.31	1.12
Achievement	**0.33**	0.78	0.90	1.07	0.81	Recognition	0.10	0.15	1.01	0.25	1.04
Ability Utilization	0.03	0.75	0.99	0.77	0.97	Co-Workers	0.04	0.14	0.99	0.18	0.94
Comfort	**0.25**	0.61	1.06	0.88	0.95	Variety	0.16	0.02	1.00	0.18	0.94
Social Status	0.08	0.59	1.16	0.68	1.10	Feedback	**0.25**	-0.06	0.86	0.16	0.87
Leisure Time	**-0.24**	0.58	1.02	0.34	1.07	Creativity	0.01	-0.01	0.94	0.00	1.01
Skill Development	0.16	0.43	1.03	0.60	0.95	Esteem	0.15	-0.12	0.95	0.02	0.88
Social Service	**0.46**	0.26	1.02	0.73	0.99	Autonomy	-0.08	-0.09	0.93	-0.16	0.95
Societal Contribution	**0.28**	0.27	1.21	0.62	1.01	Team Orientation	0.14	-0.28	1.06	-0.13	0.98
Fixed Role	**0.27**	0.26	1.01	0.53	0.96	Activity	**0.40**	-0.40	1.02	0.01	1.01
Supportive Supervision	**0.22**	0.27	1.05	0.50	1.01	Home	0.12	-0.50	1.14	-0.36	1.08
Flexible Schedule	0.05	0.23	1.04	0.29	1.05	Influence	0.19	-0.69	0.82	-0.53	0.98
Physical Development	-0.08	0.22	1.07	0.14	1.04	Travel	0.01	-0.80	1.19	-0.79	1.19
Leadership Opportunities	0.02	0.19	1.02	0.21	1.04	Independence	0.07	-0.93	1.07	-0.85	1.02

Note. $n_{Male} = 360$, $n_{Female} = 163$. d_{FM} = Effect size for Female-Male mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., Males) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

Table 13.21. WVI Scale Scores by Race/Ethnic Group

			White		Black		White Non-Hispanic		Hispanic	
Scale	d_{BW}	d_{HW}	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Advancement	-0.14	0.09	1.14	0.87	1.02	0.89	1.12	0.89	1.21	0.83
Achievement	-0.21	-0.08	0.89	0.90	0.70	0.85	0.89	0.90	0.82	0.90
Ability Utilization	-0.18	-0.18	0.77	0.96	0.59	1.00	0.79	0.98	0.61	0.90
Comfort	0.19	0.22	0.64	1.06	0.84	0.94	0.61	1.06	0.84	1.01
Social Status	-0.20	-0.04	0.67	1.14	0.44	1.09	0.68	1.14	0.63	1.12
Leisure Time	**-0.34**	**-0.38**	0.59	1.04	0.24	1.02	0.63	1.05	0.24	1.08
Skill Development	-0.07	0.16	0.48	0.97	0.41	1.20	0.45	0.98	0.61	0.89
Social Service	0.15	-0.17	0.38	0.98	0.53	1.12	0.40	0.96	0.24	1.04
Societal Contribution	-0.05	-0.21	0.40	1.16	0.34	1.18	0.44	1.16	0.19	1.06
Fixed Role	-0.01	0.00	0.34	1.04	0.33	0.96	0.35	1.05	0.35	0.95
Supportive Supervision	-0.02	0.20	0.36	1.05	0.34	0.99	0.31	1.06	0.52	0.94
Flexible Schedule	**-0.25**	0.10	0.27	1.04	0.00	1.08	0.24	1.04	0.34	1.04
Physical Development	**-0.39**	-0.03	0.28	1.03	-0.12	1.15	0.29	1.07	0.26	0.91
Leadership Opportunities	-0.06	0.03	0.19	1.01	0.13	1.06	0.17	1.02	0.21	0.99
Emotional Development	-0.10	-0.07	0.21	1.13	0.11	1.31	0.22	1.15	0.14	1.12
Recognition	-0.11	-0.01	0.21	1.04	0.09	1.00	0.22	1.00	0.20	1.13
Co-Workers	**-0.39**	-0.05	0.19	0.97	-0.18	0.90	0.20	0.98	0.15	0.86
Variety	**-0.28**	0.14	0.09	0.98	-0.18	0.91	0.08	1.00	0.22	0.89
Feedback	-0.15	-0.05	0.03	0.88	-0.11	0.90	0.04	0.91	-0.01	0.63
Creativity	-0.04	-0.11	-0.01	0.93	-0.05	1.08	0.00	0.92	-0.10	1.01
Esteem	-0.05	-0.09	-0.08	0.93	-0.13	0.96	-0.08	0.93	-0.16	0.90
Autonomy	0.16	-0.06	-0.16	0.97	0.00	0.92	-0.14	0.98	-0.20	0.95
Team Orientation	-0.02	0.01	-0.27	1.00	-0.29	1.08	-0.27	0.99	-0.25	1.06
Activity	-0.11	-0.24	-0.27	1.02	-0.38	1.01	-0.21	1.01	-0.45	1.05
Home	-0.18	-0.10	-0.43	1.12	-0.63	1.14	-0.41	1.13	-0.52	1.02
Influence	0.04	-0.04	-0.65	0.86	-0.62	0.96	-0.64	0.86	-0.67	0.89
Travel	-0.14	0.08	-0.82	1.17	-0.99	1.09	-0.81	1.18	-0.72	1.16
Independence	-0.20	-0.07	-0.92	1.02	-1.12	1.03	-0.88	1.05	-0.96	1.03

Note. n_{White} = 346. n_{Black} = 79. $n_{White\ Non\text{-}Hispanic}$ = 296. $n_{Hispanic}$ = 71. d_{BW} = Effect size for Black-White mean difference. d_{HW} = Effect size for Hispanic-White Non-Hispanic mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., White) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

Table 13.22. Intercorrelations among WVI Scale Scores

	Scale	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
1	Advancement	-																										
2	Achievement	**.37**	-																									
3	Ability Utilization	**.24**	**.35**	-																								
4	Comfort	**.31**	**.24**	**.16**	-																							
5	Social Status	**.37**	**.39**	**.17**	**.25**	-																						
6	Leisure Time	**.28**	**.17**	**.15**	**.36**	**.18**	-																					
7	Skill Development	**.26**	**.29**	**.38**	**.13**	**.18**	.02	-																				
8	Social Service	**.20**	**.31**	**.28**	**.17**	**.28**	-.01	**.28**	-																			
9	Societal Contribution	**.23**	**.39**	**.32**	.06	**.35**	-.02	**.35**	**.59**	-																		
10	Fixed Role	**.32**	**.34**	**.28**	**.22**	**.32**	**.13**	**.29**	**.31**	**.32**	-																	
11	Supportive Supervision	**.30**	**.28**	**.25**	**.30**	**.27**	**.15**	**.29**	**.34**	**.36**	**.41**	-																
12	Flexible Schedule	**.19**	**.16**	**.20**	**.44**	**.10**	**.48**	**.10**	.01	-.04	.05	**.11**	-															
13	Physical Development	**.37**	**.34**	**.24**	**.18**	**.33**	**.24**	**.31**	**.27**	**.29**	**.30**	**.27**	**.13**	-														
14	Leadership Opportunities	**.40**	**.40**	**.32**	**.16**	**.38**	**.14**	**.36**	**.45**	**.42**	**.38**	**.32**	.05	**.37**	-													
15	Emotional Development	**.25**	**.33**	**.30**	.07	**.26**	.01	**.43**	**.29**	**.38**	**.32**	**.27**	.00	**.41**	**.35**	-												
16	Recognition	**.37**	**.34**	**.25**	**.35**	**.42**	**.29**	**.14**	**.10**	**.13**	**.25**	**.23**	**.27**	**.22**	**.25**	**.16**	-											
17	Co-Workers	**.31**	**.28**	**.32**	**.29**	**.26**	**.24**	**.23**	**.33**	**.24**	**.24**	**.35**	**.29**	**.31**	**.31**	**.21**	**.31**	-										
18	Variety	**.24**	**.32**	**.26**	**.21**	**.25**	**.23**	**.20**	**.27**	**.22**	**.21**	**.20**	**.24**	**.32**	**.28**	**.16**	**.21**	**.27**	-									
19	Feedback	**.38**	**.46**	**.30**	**.28**	**.33**	**.20**	**.28**	**.23**	**.26**	**.36**	**.35**	**.14**	**.32**	**.39**	**.23**	**.47**	**.32**	**.26**	-								
20	Creativity	**.20**	**.35**	**.46**	**.27**	**.20**	**.29**	**.24**	**.19**	**.16**	**.18**	.07	**.27**	**.20**	**.32**	**.17**	**.27**	**.25**	**.34**	**.29**	-							
21	Esteem	**.30**	**.45**	**.35**	**.16**	**.37**	**.14**	**.36**	**.23**	**.34**	**.27**	**.24**	**.16**	**.26**	**.26**	**.38**	**.41**	**.31**	**.16**	**.35**	**.22**	-						
22	Autonomy	**.23**	**.20**	**.18**	**.32**	**.25**	**.37**	**.09**	**.14**	**.12**	**.13**	**.13**	**.22**	**.16**	**.24**	.08	**.16**	**.11**	**.31**	**.19**	**.33**	**.18**	-					
23	Team Orientation	**.30**	**.30**	**.32**	**.22**	**.25**	**.09**	**.33**	**.42**	**.32**	**.28**	**.34**	**.12**	**.24**	**.41**	**.29**	**.21**	**.51**	**.26**	**.32**	**.21**	**.36**	**.16**	-				
24	Activity	**.23**	**.34**	**.34**	**.18**	**.21**	.07	**.35**	**.28**	**.26**	**.37**	**.27**	**.09**	**.24**	**.24**	**.27**	**.21**	**.26**	**.38**	**.38**	**.20**	**.27**	**.17**	**.26**	-			
25	Home	**.22**	**.22**	**.27**	**.30**	**.18**	**.34**	**.20**	**.23**	**.20**	**.22**	**.17**	**.38**	**.16**	**.22**	**.21**	**.23**	**.33**	**.13**	**.23**	**.26**	**.31**	**.17**	**.25**	**.22**	-		
26	Influence	**.29**	**.40**	**.33**	**.19**	**.29**	**.14**	**.34**	**.37**	**.34**	**.33**	**.24**	**.12**	**.26**	**.50**	**.37**	**.26**	**.32**	**.30**	**.33**	**.33**	**.35**	**.28**	**.42**	**.29**	**.26**	-	
27	Travel	**.21**	**.23**	**.10**	.05	**.21**	**.10**	**.20**	.08	**.16**	**.16**	**.12**	.03	**.37**	**.26**	**.22**	**.15**	**.11**	**.33**	**.19**	**.19**	**.22**	**.19**	**.18**	**.13**	**-.17**	**.23**	-
28	Independence	**.14**	**.24**	**.19**	**.20**	**.18**	**.29**	**.10**	**.11**	**.12**	**.20**	.06	**.25**	**.26**	**.15**	**.13**	**.19**	**.09**	**.29**	**.22**	**.32**	**.21**	**.42**	.02	**.29**	**.23**	**.26**	**.23**

Note. $n = 526$. Statistically significant correlations are bolded, $p < .05$ (two-tailed).

Table 13.23. Pattern Matrix, Eigenvalues, and Communalities for WVI 6-Factor Solution

	Factors						
Scale	1	2	3	4	5	6	h^2
Ability Utilization	**0.56**	0.12	-0.02	0.02	0.06	-0.11	0.42
Skill Development	**0.53**	0.00	0.12	-0.03	0.10	0.07	0.40
Esteem	**0.43**	-0.03	0.00	*-0.40*	-0.06	0.01	0.44
Emotional Development	**0.43**	-0.15	0.12	-0.16	0.14	0.04	0.38
Activity	**0.31**	-0.02	0.02	-0.13	0.15	-0.16	0.29
Influence	**0.29**	0.05	0.12	-0.07	0.26	-0.14	0.40
Flexible Schedule	0.08	**0.63**	-0.06	-0.02	-0.17	-0.16	0.50
Co-Workers	0.19	**0.51**	0.10	-0.07	0.20	0.23	0.49
Leisure Time	-0.12	**0.50**	0.03	-0.16	-0.11	-0.28	0.48
Comfort	-0.13	**0.46**	-0.05	-0.26	0.09	-0.12	0.41
Travel	0.03	-0.06	**0.77**	-0.04	-0.12	-0.09	0.61
Home	0.28	**0.34**	*-0.38*	-0.11	0.14	-0.14	0.46
Physical Development	0.11	0.06	**0.34**	-0.20	0.12	-0.05	0.36
Variety	0.08	0.19	**0.32**	0.05	0.16	-0.27	0.36
Recognition	0.06	0.15	0.00	**-0.67**	-0.19	0.00	0.50
Social Status	-0.12	-0.05	0.08	**-0.55**	0.21	-0.09	0.43
Feedback	0.16	0.04	0.05	**-0.50**	0.04	-0.04	0.43
Advancement	0.00	0.18	0.16	**-0.43**	0.09	0.03	0.38
Achievement	0.25	-0.06	0.07	**-0.38**	0.12	-0.16	0.44
Fixed Role	0.12	-0.04	0.02	**-0.33**	0.27	-0.04	0.33
Social Service	0.03	0.01	-0.05	0.08	**0.81**	-0.06	0.63
Societal Contribution	0.16	-0.20	-0.02	-0.11	**0.59**	-0.06	0.52
Leadership Opportunities	0.12	0.02	0.19	-0.16	**0.41**	-0.06	0.46
Supportive Supervision	-0.01	0.16	0.05	-0.27	**0.36**	0.14	0.35
Team Orientation	0.24	*0.33*	0.17	0.00	**0.34**	0.24	0.48
Independence	0.10	-0.03	0.05	-0.06	-0.02	**-0.64**	0.46
Autonomy	-0.10	0.15	0.11	-0.03	0.14	**-0.52**	0.40
Creativity	*0.33*	0.19	0.08	0.05	-0.02	**-0.36**	0.39
Final Eigenvalues	7.38	1.83	1.04	0.79	0.62	0.54	

Note. h^2 = Communalities (proportion of variance accounted for in scale) by six-factor solution. Loadings for items included in the WVI composites corresponding to each factor are bolded. Cross-loadings of greater than .30 are italicized.

Based on these results, we formed six initial WVI composites by taking the average of scales that had loadings of .30 or greater on each factor.[80] For scales that had cross-loadings of .30 on multiple factors, we examined internal consistency reliability estimates and item-deleted alphas for all composites on which they could appear. From these results, we retained five scales for the Growth composite ($\alpha = .72$), five scales for the Comfort composite ($\alpha = .73$), three scales

[80] We formed these composites primarily for descriptive purposes, not necessarily for operational use. The composites provide a useful means for understanding relationships among the higher-order work value constructs that underlie the WVI. Plans for scoring the WVI operationally (in combination with the ADI) are described in Appendix I. Nevertheless, we will evaluate the potential utility of these composites within Select21 attrition database.

for the Stimulation composite (α = .61), six scales for the Status composite (α = .77), five scales for the Altruism composite (α = .77), and three scales for the Self-Direction composite (α = .63). Loadings for scales included in the final composites are bolded in Table 13.23.

Table 13.24 shows descriptives statistics and intercorrelations for the WVI composite scores. All six composites revealed good levels of variability. Recruits scored highest on the Status composite (M = 0.52) and lowest on the Self-Direction composite (M = -0.34). Examination of score distributions for the composites revealed that they were all generally normally distributed. The composites were moderately correlated (r = .27 to .63), and none were high enough to suggest that the composites failed to offer sufficient levels of unique variance.

Table 13.24. Descriptive Statistics and Intercorrelations for WVI Composite Scores

Composite		*M*	*SD*	1	2	3	4	5	6
1	Growth	0.11	0.70	(.72)					
2	Comfort	0.23	0.72	.34	(.73)				
3	Stimulation	-0.18	0.81	.46	.27	(.61)			
4	Status	0.52	0.66	.57	.47	.48	(.77)		
5	Altruism	0.22	0.77	.63	.34	.42	.59	(.77)	
6	Self-Direction	-0.34	0.74	.39	.46	.42	.41	.27	(.63)

Note. n = 526. Internal consistency reliability estimates (alpha) are shown along the diagonal in parentheses. All correlations are statistically significant, $p < .05$ (two-tailed).

WVI Fit Indices

As with the vocational interest measures, we calculated D^2 and r fit indices to assess profile similarity of recruits' WVI scale scores in relation to current and future ADI scores.[81] Before doing so, however, we rescaled SMEs' mean ADI ratings for each reinforcer onto a z-score metric so that they matched the metric of the WVI scale scores (see Appendix I).

Table 13.25 shows descriptive statistics and intercorrelations for the WVI fit indices. The mean correlation between recruits' WVI profile and the current Army ADI profile was .21 (range = -.46 to .70). A similar mean correlation was found between recruits' WVI profile and the future Army ADI profile (range = -.47 to .75). Examination of the distributions of fit index scores across recruits revealed that the distributions were positively skewed for the D^2 indices, yet generally normal for the r indices. Given the similarity of the current and future Army ADI profiles discussed earlier, it was not surprising to find that the correlations between corresponding current and future-oriented fit indices were very high (r = .95). Thus, recruits who were a good fit to the current Army also tended to be a good fit to the anticipated future Army. In line with the vocational interests results, we found that the D^2 and r-based indices were only moderately related. Again, this suggests that there were differences between ABS and ADI profiles in terms of elevation or scatter (else the D^2 and r indices would have been more highly

[81] As with the vocational interests measures, the fit indices we computed for the work values measures are primarily for descriptive purposes (although we will evaluate their effectiveness using data from the attrition database). Plans for combining person- and environment-side work values data to assess fit during the concurrent validation and for potential operational use are described in Appendix I.

correlated). Finally, no significant gender or race/ethnic differences were found on the WVI fit indices (see Tables 13.26 and 13.27).

Table 13.25. Descriptive Statistics and Intercorrelations for WVI Fit Index Scores

Fit Index	*M*	*SD*	*Skew*	1	2	3
1 WVI-Current ADI D^2	52.27	21.18	3.54	-		
2 WVI-Current ADI *r*	0.21	0.22	-0.80	-.57	-	
3 WVI-Future ADI D^2	54.05	20.11	4.23	.95	-.63	-
4 WVI-Future ADI *r*	0.17	0.23	-0.78	-.53	.95	-.66

Note. $n = 526$. All correlations are statistically significant, $p < .05$ (two-tailed).

Table 13.26. WVI Fit Index Scores by Gender

		Male		Female	
Fit Index	d_{FM}	*M*	*SD*	*M*	*SD*
WVI-Current ADI D^2	-0.04	52.53	22.02	51.68	19.44
WVI-Current ADI *r*	0.08	0.20	0.22	0.22	0.20
WVI-Future ADI D^2	-0.09	54.61	21.17	52.81	17.71
WVI-Future ADI *r*	0.04	0.16	0.23	0.17	0.22

Note. $n_{Male} = 360$, $n_{Female} = 163$. d_{FM} = Effect size for Female-Male mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., Males) are listed second in the effect size subscript. All effect sizes are nonsignificant, $p > .05$ (two-tailed).

Table 13.27. WVI Fit Index Scores by Race/Ethnic Group

			White		Black		White Non-Hispanic		Hispanic	
Fit Index	d_{BW}	d_{HW}	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
WVI-Current ADI D^2	0.11	-0.10	51.60	21.09	53.95	19.64	51.95	21.76	49.87	16.42
WVI-Current ADI *r*	-0.13	0.00	0.22	0.22	0.19	0.17	0.22	0.23	0.22	0.22
WVI-Future ADI D^2	0.19	-0.07	53.27	20.34	57.19	18.90	53.47	21.01	51.91	15.64
WVI-Future ADI *r*	-0.16	-0.02	0.18	0.24	0.14	0.18	0.17	0.24	0.17	0.23

Note. $n_{White} = 346$. $n_{Black} = 79$. $n_{White\ Non\text{-}Hispanic} = 296$. $n_{Hispanic} = 71$. d_{BW} = Effect size for Black-White mean difference. d_{HW} = Effect size for Hispanic-White Non-Hispanic mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., White) are listed second in the effect size subscript. All effect sizes are nonsignificant, $p > .05$ (two-tailed).

Next Steps for the WVI

Few modifications are proposed for the content or structure of the WVI for the concurrent validation. The changes planned for the validation effort center on remedying the programming bugs described earlier. Nevertheless, we leave open the possibility of replacing some of the reinforcers on the current WVI with other reinforcers included in the pilot and faking research versions of the WVI (e.g., Security, Compensation). As the Select21 attrition database matures, we will have the ability to examine relationships between each WVI reinforcer and attrition at various stages of Soldiers' first-term. If these findings suggest that preferences for Security and Compensation predict attrition, whereas other current reinforcers are not, we will consider substituting Security and Compensation into the WVI for reinforcers that lack predictive validity.

Furthermore, as with the vocational interests predictor measures, we propose not calculating WVI fit indices based on the future Army ADI profile in subsequent Select21 research (e.g., because of the high correlations between current and future Army ADI profiles and the overlap between WVI fit indices based on these two sets of data). We will, however, use both the current and future ADI data to model relations between the WVI, ADI, and criteria of interest in subsequent efforts.

Lastly, at some point we would also like to gather test-retest data on the WVI to assess (a) the consistency of individuals' preference for each reinforcer across occasions and (b) the average consistency with which reinforcers are rank ordered by individuals across occasions. Given the rank-order format of the WVI, use of traditional internal consistency indices to assess the reliability of the WVI scale scores is problematic (Nunnally & Bernstein, 1994). Thus, we will continue to explore options for obtaining samples that would allow us to assess test-retest reliability for the WVI scores.

Person-Side Expectations Measure: Army Beliefs Survey

Description of Measure

The ABS is a 30-item measure designed to assess recruits' expectations regarding the extent to which the Army supports the reinforcers covered by the WVI.[82] Respondents are asked to assign each reinforcer (e.g., Soldiers learn new skills) to one of three categories that indicates whether they think the statement describes an experience that (a) few Soldiers will have during their first enlistment, (b) some Soldiers will have during their first enlistment, or (c) most Soldiers will have during their first enlistment. Respondents are instructed to think of what experiences they expect Soldiers to have once they have complete initial entry training and are assigned to their unit. ABS data will be used to determine the extent to which the expectations of recruits are consistent with what reinforcers the Army actually supplies (assessed via the ADI). We also will use ABS data to determine whether expectations moderate relations between needs-supplies fit and various P-E fit criteria.

The ABS was administered to recruits during pilot testing and in the faking research (honest condition only). No substantive changes were made to the instrument based on analyses of these data and qualitative feedback from respondents.

Field Test Results

Sample

The ABS was administered to 225 recruits during the predictor field test. As with the vocational interests expectations measure (i.e., the PSES), recruits completed the ABS as time permitted. As such, over half the field test participants did not complete the ABS. Of the recruits with ABS data, 29 were removed from the sample because they failed to provide complete data.

[82] Because administration time was less of an issue for the ABS than for the WVI, we included two of the reinforcers dropped from the field test version of the WVI, namely Security and Compensation.

Complete data on the ABS were needed to calculate ABS fit indices for each recruit. Predictor field test problem logs revealed no issues with the remaining 196 recruits' ABS data.

Descriptive Statistics

Table 13.28 shows descriptive statistics for the ABS scale scores. On average, recruits had the highest expectations regarding Physical Development, Emotional Development, Skill Development and Team Orientation, and the lowest expectations regarding Independence, Autonomy, Home, Flexible Schedule, and Creativity. Several of the ABS scale scores were highly negatively skewed. The scales with the highest negative skews tended to be those that are in high supply in the Army based on the ADI data (e.g., Physical, Emotional, and Skill Development). Such findings indicate that recruits generally had a good sense for what reinforcers the Army supplies. Interestingly, few highly positively skewed distributions were found, even on reinforcers that are in low supply in the Army. Thus, although it seems recruits have a sense for what reinforcers the Army supplies, they appear to have a poorer sense for what reinforcers the Army *does not* supply.

Table 13.28. Descriptive Statistics for ABS Scale Scores

Scale	*M*	*SD*	*Skew*	Scale	*M*	*SD*	*Skew*
Physical Development	2.91	0.35	-4.13	Travel	2.48	0.63	-0.82
Emotional Development	2.87	0.38	-3.07	Esteem	2.47	0.66	-0.86
Skill Development	2.84	0.44	-2.89	Social Status	2.45	0.65	-0.79
Team Orientation	2.83	0.41	-2.19	Influence	2.37	0.69	-0.64
Fixed Role	2.80	0.46	-2.28	Recognition	2.36	0.68	-0.59
Co-Workers	2.79	0.43	-1.82	Leadership Opportunities	2.34	0.62	-0.40
Activity	2.73	0.51	-1.76	Variety	2.21	0.67	-0.28
Social Service	2.71	0.53	-1.64	Leisure Time	2.19	0.74	-0.32
Achievement	2.68	0.58	-1.63	Comfort	1.81	0.67	0.24
Societal Contribution	2.63	0.61	-1.41	Creativity	1.76	0.69	0.35
Advancement	2.60	0.59	-1.15	Home	1.73	0.72	0.44
Ability Utilization	2.57	0.55	-0.77	Flexible Schedule	1.71	0.75	0.52
Supportive Supervision	2.52	0.61	-0.89	Autonomy	1.58	0.63	0.63
Feedback	2.50	0.69	-1.04	Independence	1.43	0.64	1.19

Note. $n = 196$. Scales are shown in descending order by mean score.

Subgroup differences on the ABS scales. Tables 13.29 and 13.30 show mean ABS scale scores by gender and race/ethnicity, respectively. Statistically significant gender differences were found for only two of the 28 ABS scales. On average, females viewed the Army as providing more opportunities for Skill Development than males ($d = 0.30$). Conversely, males expected the Army to provide more Feedback on performance than females ($d = -0.33$). Only two of the ABS scales demonstrated significant differences with regard to race/ethnicity. Whites expected the Army to provide more Feedback on performance than Blacks ($d = -0.52$), whereas Blacks expected the Army to provide more opportunities for Variety in their work ($d = 0.48$).

Table 13.29. ABS Scale Scores by Gender

Scale	d_{FM}	Male		Female		Scale	d_{FM}	Male		Female	
		M	*SD*	*M*	*SD*			*M*	*SD*	*M*	*SD*
Physical Development	-0.07	2.91	0.33	2.89	0.42	Travel	-0.06	2.49	0.65	2.45	0.57
Emotional Development	0.00	2.87	0.38	2.87	0.39	Esteem	0.09	2.45	0.68	2.51	0.60
Skill Development	**0.30**	2.80	0.48	2.95	0.30	Social Status	-0.20	2.49	0.64	2.36	0.68
Team Orientation	-0.09	2.84	0.39	2.80	0.45	Influence	-0.01	2.37	0.66	2.36	0.75
Fixed Role	-0.12	2.81	0.43	2.76	0.54	Recognition	-0.08	2.38	0.66	2.33	0.72
Co-Workers	-0.03	2.79	0.44	2.78	0.42	Leadership Opportunities	-0.08	2.36	0.62	2.31	0.63
Activity	-0.13	2.76	0.49	2.69	0.54	Variety	0.03	2.20	0.65	2.22	0.74
Social Service	0.09	2.70	0.53	2.75	0.52	Leisure Time	-0.02	2.20	0.74	2.18	0.75
Achievement	-0.05	2.69	0.58	2.65	0.58	Comfort	0.06	1.79	0.68	1.84	0.63
Societal Contribution	0.04	2.63	0.62	2.65	0.55	Creativity	0.04	1.76	0.69	1.78	0.69
Advancement	-0.03	2.60	0.60	2.58	0.57	Home	0.20	1.69	0.72	1.84	0.71
Ability Utilization	-0.12	2.59	0.55	2.53	0.54	Flexible Schedule	0.22	1.67	0.74	1.84	0.76
Supportive Supervision	0.01	2.52	0.61	2.53	0.63	Autonomy	-0.24	1.62	0.63	1.47	0.63
Feedback	**-0.33**	2.56	0.67	2.35	0.73	Independence	-0.12	1.46	0.64	1.38	0.65

Note. $n_{Male} = 140$, $n_{Female} = 55$. d_{FM} = Effect size for Female-Male mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., Males) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

Table 13.30. ABS Scale Scores by Race/Ethnic Group

			White		Black		White Non-Hispanic		Hispanic	
Scale	d_{BW}	d_{HW}	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Physical Development	-0.07	0.04	2.92	0.32	2.90	0.31	2.92	0.33	2.93	0.26
Emotional Development	-0.09	-0.48	2.87	0.40	2.83	0.38	2.91	0.32	2.76	0.58
Skill Development	-0.26	0.21	2.84	0.42	2.73	0.58	2.84	0.44	2.93	0.26
Team Orientation	-0.35	-0.05	2.84	0.39	2.70	0.53	2.85	0.39	2.83	0.38
Fixed Role	-0.55	-0.36	2.83	0.42	2.60	0.62	2.84	0.42	2.69	0.60
Co-Workers	-0.18	-0.37	2.81	0.40	2.73	0.58	2.83	0.38	2.69	0.47
Activity	0.09	0.13	2.72	0.53	2.77	0.43	2.72	0.54	2.79	0.41
Social Service	-0.06	-0.22	2.70	0.52	2.67	0.61	2.73	0.49	2.62	0.62
Achievement	-0.14	-0.43	2.68	0.57	2.60	0.62	2.72	0.56	2.48	0.63
Societal Contribution	-0.24	0.06	2.67	0.56	2.53	0.68	2.66	0.56	2.69	0.54
Advancement	-0.16	-0.39	2.62	0.55	2.53	0.68	2.68	0.52	2.48	0.57
Ability Utilization	-0.07	-0.11	2.57	0.56	2.53	0.51	2.58	0.56	2.52	0.51
Supportive Supervision	0.27	0.13	2.47	0.63	2.63	0.56	2.47	0.63	2.55	0.57
Feedback	**-0.52**	-0.32	2.55	0.67	2.20	0.81	2.59	0.64	2.38	0.73
Travel	-0.07	-0.25	2.54	0.60	2.50	0.57	2.56	0.58	2.41	0.73
Esteem	-0.15	-0.03	2.46	0.63	2.37	0.81	2.47	0.64	2.45	0.51
Social Status	-0.21	-0.32	2.50	0.63	2.37	0.72	2.53	0.58	2.34	0.77
Influence	-0.01	-0.06	2.31	0.68	2.30	0.75	2.32	0.67	2.28	0.75
Recognition	-0.17	-0.06	2.38	0.66	2.27	0.74	2.39	0.68	2.34	0.55
Leadership Opportunities	0.02	-0.33	2.36	0.62	2.37	0.56	2.41	0.59	2.21	0.73
Variety	**0.48**	0.14	2.12	0.66	2.43	0.63	2.14	0.67	2.24	0.64
Leisure Time	-0.06	0.05	2.24	0.72	2.20	0.76	2.24	0.72	2.28	0.75
Comfort	-0.21	0.09	1.84	0.66	1.70	0.65	1.84	0.65	1.90	0.72
Creativity	0.00	-0.19	1.77	0.70	1.77	0.68	1.78	0.69	1.66	0.67
Home	0.08	0.14	1.74	0.74	1.80	0.71	1.72	0.74	1.83	0.71
Flexible Schedule	0.16	0.12	1.67	0.76	1.80	0.76	1.67	0.74	1.76	0.79
Autonomy	0.11	-0.11	1.57	0.60	1.63	0.67	1.59	0.59	1.52	0.69
Independence	0.31	-0.15	1.38	0.60	1.57	0.68	1.41	0.62	1.31	0.60

Note. n_{White} = 129. n_{Black} = 30. $n_{White\ Non\text{-}Hispanic}$ = 111. $n_{Hispanic}$ = 29. d_{BW} = Effect size for Black-White mean difference. d_{HW} = Effect size for Hispanic-White Non-Hispanic mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., White) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

Scale Intercorrelations and Factor Structure

Table 13.31 shows intercorrelations among the ABS scale scores. On average, the ABS scales were only modestly correlated (mean $r = .14$). EFA was used to examine the factor structure underlying the 28 ABS scales. In the initial analysis, we did not specify the number of factors to be extracted. Results indicated that eight factors had eigenvalues over 1.0. The scree plot indicated a distinct drop off in the magnitude of eigenvalues after the sixth factor. In light of these results, we conducted a follow-up EFA that constrained the solution to six factors. The pattern matrix for this 6-factor solution revealed several cross loadings, and as such we

conducted two more follow-up EFAs that constrained the solution to four and five factors, respectively. Comparing the results from these analyses, we adopted the 4-factor solution given the interpretability of its factors and few cross loadings. The pattern matrix resulting from this model, along with final eigenvalues and communalities, are shown in Table 13.32.

The factors underlying the ABS scales appear to reflect three expectations-related constructs: (a) expectations regarding reinforcers that are generally supplied by the Army (Supported Reinforcers), (b) expectations regarding reinforcers that are generally unsupported by the Army (Unsupported Reinforcers), and (c) expectations that the Army will provide opportunities for recognition and achievement (Recognition/Achievement). The naming of the first two factors reflects the fact that (a) the majority of ABS scales in the Supported Reinforcers composite reflect reinforcers in high supply in the Army (based on ADI data), and (b) the majority of the ABS scales in the Unsupported Reinforcers composite reflect reinforcers in low supply in the Army. The naming of the third factor followed directly from its content. The fourth factor was difficult to interpret and the three scales that loaded on it (Variety, Influence, Leadership Opportunities) resulted in a low level of internal consistency when combined (α = .53). As such, we did not consider the fourth factor for subsequent analyses.

Based on these results, we formed three ABS composites by taking the mean of scales that had loadings of .30 or greater on each of the first three factors. Because the Leisure Time scale cross-loaded on the second and fourth factors, we examined internal consistency reliability of the Unsupported Reinforcers composite with and without this scale. We retained nine scales for the Supported Reinforcers composite (α = .73), seven scales for the Unsupported Reinforcers composite (α = .73, including the Leisure Time scale), and six scales for the Recognition/Achievement composite (α = .73). Loadings for scales included in each composite are bolded in Table 13.32.

Table 13.33 shows descriptive statistics and intercorrelations for the ABS composite scores. All three composites revealed good levels of variability. Recruits scored highest on the Supported Reinforcers composite (M = 2.80) and lowest on the Unsupported Reinforcers composite (M = 1.74). Examination of score distributions for the composites revealed that the Supported Reinforcers and Recognition/Achievement composites were negatively skewed, whereas the Unsupported Reinforcers composite was slightly positively skewed. Interestingly, the Supported Reinforcers and Unsupported Reinforcers composites were uncorrelated (r = .02). This suggests that the accuracy of recruits' expectations regarding reinforcers supported by the Army has little relation to the accuracy of recruits' expectations regarding unsupported reinforcers. Like the scale-level results presented above, these results suggest that recruits have a better idea of what the Army supplies than what it does not supply.

Table 13.31. Intercorrelations among ABS Scale Scores

	Scale	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
1	Physical Development	-																										
2	Emotional Development	**.33**	-																									
3	Skill Development	**.43**	**.34**	-																								
4	Team Orientation	**.17**	**.16**	**.22**	-																							
5	Fixed Role	**.33**	**.24**	**.35**	**.34**	-																						
6	Co-Workers	**.18**	**.28**	**.20**	**.23**	**.25**	-																					
7	Activity	**.21**	**.28**	**.22**	.00	**.23**	.10	-																				
8	Social Service	**.27**	.12	**.20**	**.15**	**.35**	**.27**	**.28**	-																			
9	Achievement	**.21**	**.14**	**.20**	**.18**	**.36**	**.16**	.08	**.26**	-																		
10	Societal Contribution	**.22**	**.28**	**.26**	**.15**	**.30**	**.15**	**.19**	**.32**	**.18**	-																	
11	Advancement	.09	**.21**	**.19**	.01	**.20**	**.25**	.09	.13	**.33**	**.15**	-																
12	Ability Utilization	**.19**	**.18**	**.23**	.10	**.25**	**.25**	**.18**	**.19**	**0.3**	**.24**	**.29**	-															
13	Supportive Supervision	**.15**	.09	**.27**	**.18**	**.17**	.06	.13	.09	**.19**	.10	**.19**	**.26**	-														
14	Feedback	.13	.07	**.24**	**.18**	**.25**	**.20**	.13	**.15**	**.32**	.13	**.30**	**.23**	**.21**	-													
15	Travel	.11	**.18**	.06	.07	**.14**	**.17**	**.23**	**.15**	**.15**	**.18**	**.14**	**.18**	-.06	.08	-												
16	Esteem	**.19**	**.18**	**.22**	.04	**.17**	.09	.08	**.17**	**.24**	**.32**	**.31**	**.42**	**.29**	**.16**	.08	-											
17	Social Status	.09	.11	**.22**	.13	**.27**	.14	.06	**.24**	**.21**	**.30**	**.20**	.12	.06	**.19**	**.21**	**.17**	-										
18	Influence	**.14**	.07	**.18**	**.18**	**.24**	**.18**	.11	**.19**	**.24**	**.16**	.11	**.37**	.11	**.17**	.11	**.29**	**.22**	-									
19	Recognition	-.01	.08	.12	-.01	**.15**	.12	**.18**	**.25**	**.35**	**.17**	**.37**	**.33**	**.20**	**.39**	**.18**	**.34**	**.24**	**.20**	-								
20	Leadership Opportunities	.05	-.03	.12	.03	.10	.11	.08	**.18**	**.24**	.11	**.25**	**.31**	**.20**	**.15**	.14	**.19**	**.25**	**.31**	**.24**	-							
21	Variety	-.07	.00	.04	.08	.09	.12	**.16**	**.16**	.02	.05	.07	.12	**.17**	.05	**.16**	**.15**	.09	**.24**	**.20**	**.30**	-						
22	Leisure Time	-.03	-.02	.14	-.01	.02	-.04	-.11	-.06	.05	.12	**.23**	**.18**	.05	.11	.07	.10	**.16**	.06	.10	.08	-.04	-					
23	Comfort	-.05	-.06	.07	-.01	.04	.04	-.05	-.03	.04	.12	**.19**	**.22**	.09	.13	.08	**.24**	.07	.05	.09	**.17**	.06	**.38**	-				
24	Creativity	.06	-.08	**.14**	.05	.11	**.16**	-.05	.10	**.14**	.03	**.18**	**.29**	**.15**	**.19**	.02	**.33**	.13	**.34**	**.27**	**.37**	**.18**	**.28**	**.35**	-			
25	Home	-.02	-.05	.00	.05	-.07	.07	-.11	-.02	.00	.08	.09	.11	.09	.00	.05	**.22**	.05	.11	.06	**.19**	.08	.13	**.28**	**.29**	-		
26	Flexible Schedule	-.04	-.07	.05	-.05	-.06	.08	.07	.09	.05	.05	.11	**.16**	.02	.07	.03	**.20**	.03	**.20**	**.22**	**.23**	**.21**	**.27**	**.35**	**.52**	**.35**	-	
27	Autonomy	-.08	-.08	-.11	-.05	-.08	.03	-.05	.03	-.02	.09	**.19**	**.17**	.04	-.01	.03	.12	**.17**	.07	.02	**.16**	.03	**.16**	**.32**	**.25**	.10	.12	-
28	Independence	-.12	-.05	-.10	-.02	.02	.03	.04	-.05	-.04	.05	.00	.04	-.11	.06	.01	**.15**	.09	.13	.03	**.20**	**.16**	**.15**	**.33**	**.34**	**.24**	**.35**	**.38**

Note. $n = 196$. Statistically significant correlations are bolded, $p < .05$ (two-tailed).

Table 13.32. Pattern Matrix, Eigenvalues, and Communalities for ABS 4-Factor Solution

Scale	ADI Supply Category	Factors 1	2	3	4	h^2
Physical Development	High	**0.65**	-0.05	0.08	-0.08	0.37
Skill Development	High	**0.59**	0.03	-0.09	-0.13	0.39
Fixed Role	Mid	**0.56**	-0.06	-0.11	0.06	0.40
Emotional Development	High	**0.54**	-0.10	0.01	-0.08	0.29
Societal Contribution	Mid	**0.46**	0.11	-0.06	-0.02	0.26
Team Orientation	High	**0.39**	0.00	0.06	0.04	0.15
Social Service	High	**0.34**	-0.06	-0.13	0.28	0.30
Co-Workers	High	**0.34**	0.05	-0.07	0.12	0.18
Activity	Mid	**0.30**	-0.11	-0.04	0.23	0.19
Social Status	Mid	0.21	0.10	-0.20	0.06	0.16
Travel	Mid	0.17	0.02	-0.10	0.14	0.09
Comfort	Low	0.03	**0.63**	-0.07	-0.24	0.45
Creativity	Low	0.01	**0.62**	-0.14	0.14	0.50
Independence	Low	-0.01	**0.60**	0.20	0.15	0.36
Flexible Schedule	Low	-0.06	**0.59**	-0.01	0.17	0.40
Home	Low	0.01	**0.46**	0.05	0.03	0.20
Autonomy	Low	-0.06	**0.42**	-0.02	-0.04	0.18
Leisure Time	Mid	0.01	**0.42**	-0.18	*-0.34*	0.32
Recognition	Mid	-0.22	-0.07	**-0.78**	0.19	0.55
Advancement	High	0.02	0.07	**-0.59**	-0.15	0.38
Achievement	High	0.14	-0.10	**-0.50**	0.03	0.33
Feedback	High	0.07	-0.02	**-0.48**	-0.03	0.26
Ability Utilization	Mid	0.22	0.20	**-0.35**	0.05	0.35
Esteem	Mid	0.19	0.27	**-0.31**	0.01	0.31
Supportive Supervision	Mid	0.13	0.02	-0.29	0.00	0.14
Variety	Low	-0.03	0.14	-0.06	**0.48**	0.27
Influence	Low	0.20	0.21	-0.12	**0.32**	0.29
Leadership Opportunities	Mid	-0.02	0.28	-0.25	**0.31**	0.32
Final Eigenvalues		4.46	2.30	0.84	0.77	

Note. ADI Supply Category = supply category for each ABS reinforcer (based on ADI data). Loadings for items included in ABS composites corresponding to each factor are bolded. Cross-loadings greater than .30 are italicized.

Table 13.33. Descriptive Statistics and Intercorrelations for ABS Composite Scores

Composite	*M*	*SD*	*Skew*	1	2	3
1 Supported Reinforcers	2.80	0.26	-1.65	(.73)		
2 Unsupported Reinforcers	1.74	0.43	0.62	.02	(.73)	
3 Recognition/Achievement	2.53	0.41	-0.99	**.45**	**.30**	(.73)

Note. $n = 196$. Internal consistency reliability estimates (alpha) are shown along the diagonal in parentheses Statistically significant correlations are bolded, $p < .05$ (one-tailed).

Relations between Expectations and Work Values

Given the ABS and WVI are based on the same 28 reinforcers, an important question to address is whether recruits' expectations can be differentiated from their work values. In other words, are recruits' expectations regarding the Army's provision of the 28 reinforcers distinct from their preferences for the 28 reinforcers? To answer this question, we examined correlations between recruits' scores on corresponding ABS and WVI scales. The correlations presented in Table 13.34 reveal that recruits' expectations and preferences were generally unrelated (mean r = .05). Only six of 28 scales showed significant correlations, and in these cases, the correlations were relatively small (r = .14 to .27). These results (which are consistent with the interests-expectations results discussed earlier) indicate that recruits' expectations and preferences with regard to the reinforcers assessed on the ABS and WVI are distinct.

Table 13.34. Correlations between Corresponding ABS and WVI Scale Scores

Scale	*r*	Scale	*r*
Social Status	**.16**	Travel	.13
Advancement	-.13	Physical Development	.10
Autonomy	-.10	Ability Utilization	.02
Supportive Supervision	.00	Creativity	.03
Leisure Time	**.15**	Recognition	**.27**
Comfort	.08	Co-Workers	**.14**
Achievement	.05	Activity	-.01
Societal Contribution	**.17**	Flexible Schedule	.01
Independence	-.07	Skill Development	.02
Social Service	**.18**	Home	.02
Fixed Role	.08	Esteem	.02
Variety	.09	Emotional Development	.06
Leadership Opportunities	.12	Influence	.00
Feedback	.08	Team Orientation	-.02

Note. n = 147. Statistically significant correlations are bolded, $p < .05$ (one-tailed).

ABS Fit Indices

As with previous instruments, we calculated D^2 and r fit indices to assess profile similarity of recruits' ABS scale scores in relation to current ADI scores. However, before doing so, we rescaled the ADI ratings from a 5- to a 3-point metric (in the same way we rescaled the AES and FAES ratings for computing the IFQ fit indices) so they matched the metric of the ABS. We then recalculated mean scores for each ADI reinforcer for calculating the fit indices. Table 13.35 shows descriptive statistics and intercorrelations for the ABS fit indices. The mean correlation between recruits' ABS profile and the current Army ADI profile was .52 (range = -.20 to .83). Score distributions for the fit indices were positively skewed for D^2 and negatively skewed for the r. No significant gender or race/ethnic differences were found for these fit indices (see Tables 13.36 and 13.37).

Table 13.35. Descriptive Statistics and Intercorrelations for ABS Fit Index Scores

Fit Index	M	SD	$Skew$	r
1 ABS-Current ADI D^2	12.06	5.50	1.17	-
2 ABS-Current ADI r	0.52	0.22	-0.96	-.77

Note. n = 193-196. All correlations are statistically significant, $p < .05$ (one-tailed).

Table 13.36. ABS Fit Index Scores by Gender

		Male		Female	
Fit Index	d_{FM}	M	SD	M	SD
ABS-Current ADI D^2	0.10	11.89	5.65	12.43	5.20
ABS-Current ADI r	-0.12	0.53	0.22	0.50	0.20

Note. n_{Male} = 137-140, n_{Female} = 55. d_{FM} = Effect size for Female-Male mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., Males) are listed second in the effect size subscript. All effect sizes were nonsignificant, $p < .05$ (two-tailed).

Table 13.37. ABS Fit Index Scores by Race/Ethnic Group

			White		Black		White Non-Hispanic		Hispanic	
Fit Index	d_{BW}	d_{HW}	M	SD	M	SD	M	SD	M	SD
ABS-Current ADI D^2	0.48	0.36	11.39	5.02	13.79	6.94	11.02	4.99	12.82	4.48
ABS-Current ADI r	-0.52	-0.39	0.55	0.20	0.44	0.28	0.56	0.20	0.48	0.20

Note. n_{White} = 127-129. n_{Black} = 29-30. $n_{White\ Non\text{-}Hispanic}$ = 109-111. $n_{Hispanic}$ = 29. d_{BW} = Effect size for Black-White mean difference. d_{HW} = Effect size for Hispanic-White Non-Hispanic mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., White) are listed second in the effect size subscript. All effect sizes were nonsignificant, $p < .05$ (two-tailed).

Next Steps for the ABS

As with the other P-E fit expectations measures, the ABS will not be administered in the concurrent validation. Again, because Soldiers participating in the validation effort will have 18-36 months of service, it would not be appropriate to ask them about their expectations prior to entering the Army. Nevertheless, we will include ABS data gathered during the pilot, faking research, and predictor field data collections in the Select21 attrition database to examine relations between the ABS (e.g., scale scores, composites, fit indices) and attrition. As was with the WVI fit indices, we plan to examine only fit indices based on the current Army ADI profile, but will consider current and future ADI data for modeling relations between the ABS, ADI, and relevant criteria in subsequent Select21 efforts.

TEMPERAMENT MEASURES

We now discuss measures designed to assess fit with regard to temperament. We begin with the environment-side measure.

Environment-Side Demands Measure: Work Styles Supply Survey

Description of Measure

The Work Styles Supply Survey (WSSS) is a rank-order instrument designed to assess the temperament-related requirements of work performed by first-term Soldiers. The WSSS parallels the WSI (see Chapter 9) in both content and format. In fact, the only difference between these measures is that instead of asking recruits to rank order 16 types of work in terms of how effectively they think they would perform them, the WSSS asks Army SMEs (i.e., NCOs) to rank order the types of work in terms of how well they describe work performed by first-term Soldiers. Like the WSI, each type of work described on the WSSS is linked to a different dimension of temperament (e.g., Dependability, Concern for Others). The WSSS provides the environment-side data against which to compare recruits' (a) Work Suitability Inventory (WSI) scores to assess abilities-demands fit, and (b) Army Work Knowledge Survey (AWKS) scores to assess expectations-demands fit with regard to temperament.

Field Test Results

Sample

The WSSS was administered to 100 NCOs who provided performance ratings to their subordinates in the criterion field test (see Chapter 3). As with the other environment-side measures discussed in this chapter, we first examined the data for individual NCOs who had ratings that were very inconsistent with the others. Specifically, we calculated rater-total correlations (i.e., correlations between each NCO's WSSS profile and the mean WSSS profile across the other NCOs), and removed data from NCOs who had a profile that correlated less than .10 with the mean profile of the other NCOs. The ratings of 12 NCOs failed to meet this criterion, and thus were removed from the sample.

Descriptive Statistics

Table 13.38 displays descriptive statistics for the WSSS dimensions. Scale scores for each dimension were created by (a) taking the ranking given to the dimension by each NCO and subtracting it from 17, and (b) averaging the resulting values across NCOs. Thus, dimension scores range from 1 to 16, with higher scores being indicative of types of work that NCOs deemed most descriptive of work performed by first-term Soldiers.

NCOs indicated that types of work requiring Attention to Detail, Social Orientation, Dependability, and Adaptability/Flexibility were most descriptive of work performed by first-term Soldiers. Conversely, NCOs indicated that types of work requiring Concern for Others, Innovation, and Independence were least descriptive of work performed by first-term Soldiers. ICCs were used to estimate the consistency with which NCOs rank-ordered the dimensions. The resulting single- and k-rater reliability estimates were .21 and .96, respectively. Although the

single-rater reliability suggests that there was little consistency between the ratings of any two SMEs, the reliability of the mean profile (based on 88 NCOs) was very high.

Table 13.38. Descriptive Statistics for WSSS Scale Scores

Scale	*M*	*SD*	SE_M
Attention to Detail	11.94	3.93	0.42
Social Orientation	11.67	3.79	0.40
Dependability	10.90	3.98	0.42
Adaptability/Flexibility	10.78	4.20	0.45
Stress Tolerance	9.67	4.25	0.45
Self-Control	9.47	3.50	0.37
Cultural Tolerance	9.43	4.77	0.51
Energy	8.58	4.15	0.44
Achievement/Effort	8.49	4.03	0.43
Cooperation	7.92	3.90	0.42
Initiative	7.23	4.18	0.45
Persistence	7.18	4.39	0.47
Leadership Orientation	6.73	4.87	0.52
Independence	5.92	4.16	0.44
Innovation	5.08	3.23	0.34
Concern for Others	5.01	4.05	0.43

Note. $n = 88$. SE_M = standard error of the mean. Scales are presented in descending order by mean score.

As for interrater agreement on individual dimensions, *SD*s were generally quite sizable. For comparison, the *SD*s for uniformly and normally distributed ratings on a 16-point scale would be 4.61 and 3.45, respectively. Because most of the observed *SD*s were in between these values (mean $SD = 4.09$), it suggests raters tended to disagree with regard to how well each dimension describes the work performed by first-term Soldiers (James et al., 1984).

Next Steps for the WSSS

As with the other environment-side measures, we will not collect further data on the WSSS for Select21. The purpose of the WSSS was to create an Army temperament profile against which recruits' ratings on the WSI and AWKS could be compared to assess abilities-demands and expectations-demands fit with regard to temperament, respectively. Sufficient data was obtained during the criterion field test to create the WSSS profile for use in subsequent efforts. Nevertheless, the mean WSSS profile should be used cautiously given the high level of disagreement found at the dimension level.

Person-Side Abilities Measure: Work Suitability Inventory

Description of Measure

The WSI is a computerized rank-order instrument designed to assess the degree to which the recruits feel they can perform 16 different types of work. Each type of work is linked to a different dimension of temperament. Further details on the format, content, development, and data gathered on the WSI are provided in Chapter 9. In the section that follows, we discuss

results of analyses that assessed the similarity between recruits' WSI profiles and the mean WSSS profile provided by NCOs. The WSI provides the person-side abilities measure for assessing abilities-demands fit with regard to temperament.

Field Test Results

As noted in Chapter 9, usable WSI data were available for 630 recruits who participated in the predictor field test. For each recruit, we calculated a Spearman correlation coefficient to index the similarity of his/her WSI profile to the mean WSSS profile. Prior to calculating the Spearman *r*s, we converted the WSI and WSSS data to ranks. For the WSI, we subtracted the rank recruits gave to each dimension from 17; thus, higher ranked dimensions had higher scores. For the WSSS data, we rank ordered the means from lowest to highest; thus, like the WSI, higher ranked dimensions had higher scores. We then calculated the Spearman r between each recruits' WSI and WSSS profiles. The mean correlation between these profiles was quite low ($M = .04$, $SD = .28$, range = -.64 to .76). In addition, correlations across recruits were normally distributed. These findings indicate that while on average recruits' temperament profiles seemed to be out of line with the temperament profile characteristic of Army work, there was a good deal of variation in similarity across recruits.

As for subgroup effects, no statistically significant differences were found on the WSI profile correlations for gender (Males: $M = 0.05$, $SD = 0.29$; Females: $M = 0.01$, $SD = 0.25$) or race/ethnicity (Whites: $M = 0.04$, $SD = 0.28$; Blacks: $M = 0.03$, $SD = 0.28$; Hispanics: $M = 0.05$, $SD = 0.24$).[83]

Next Steps for the WSI

Unlike the other person-side measures discussed in this chapter, the WSI will primarily be used as a stand-alone measure during the remainder of this project. However, we will examine the potential utility of the using the WSI-WSSS profile correlations to predict criteria as such data become available.

Person-Side Expectations Measure: Army Work Knowledge Survey

Description of Measure

The AWKS is a 16-item instrument designed to assess recruits' expectations regarding the temperament-related requirements of Army work. As with the WSI and WSSS, each item describes a type of work that is linked to a temperament dimension. Although the item content of the AWKS is completely parallel to that of the WSI and WSSS, the format and instruction set are different. Respondents are asked to indicate the degree to which they agree that each type of work is required of first-term Soldiers using a 5-point Likert-type scale ranging from *strongly disagree* (1) to *strongly agree* (5). The AWKS provides the person-side expectations measure for the assessment of expectations-demands fit with regard to temperament.

---

[83] Sample sizes for subgroup analyses were as follows: $n_{\text{Males}} = 437$; $n_{\text{Females}} = 189$; $n_{\text{Whites}} = 413$; $n_{\text{Blacks}} = 94$; $n_{\text{White Non-Hispanics}} = 357$; $n_{\text{Hispanics}} = 83$.

Field Test Results

Sample

The AWKS was administered to 255 recruits as part of the predictor field test (on a "as time permitted basis" like the other expectations measures). Of the recruits with AWKS data, 10 were removed from the sample because they failed to provide complete data, which were needed to calculate a fit index for each recruit. Predictor field test problem logs revealed no issues with the remaining recruits' AWKS data.

Descriptive Statistics

Table 13.39 shows descriptive statistics for each AWKS dimension. Recruits tended to indicate that nearly all the types of work examined on the AWKS were required of first-term Soldiers (average $M = 3.98$). This was also evidenced by the finding that 10 of the 16 AWKS dimensions had response distributions that were highly negatively skewed (i.e., *Skew* < -1.0).

Table 13.39. Descriptive Statistics for AWKS Scale Scores

Scale	*M*	*SD*	*Skew*
Dependability	4.30	0.82	-1.24
Attention to Detail	4.29	0.83	-1.19
Energy	4.23	0.91	-1.38
Cultural Tolerance	4.23	0.89	-1.32
Social Orientation	4.22	0.91	-1.33
Self-Control	4.21	0.88	-1.27
Adaptability/Flexibility	4.14	0.81	-1.10
Cooperation	4.09	0.88	-1.35
Stress Tolerance	4.08	0.98	-1.09
Achievement/Effort	4.06	0.95	-1.19
Persistence	4.04	0.94	-0.84
Initiative	3.96	0.98	-0.82
Leadership Orientation	3.79	1.01	-0.68
Concern for Others	3.56	1.04	-0.53
Innovation	3.42	1.07	-0.24
Independence	3.13	1.19	0.03

Note. $n = 245$. Dimensions are presented in descending order by mean score.

Subgroup differences on the AWKS scales. Tables 13.40 and 13.41 show mean AWKS scale scores by gender and race/ethnicity, respectively. Statistically significant gender differences were found for six of the 16 AWKS scales. In all cases, females had significantly higher expectations than males. Nonetheless, the magnitudes of these effects sizes were relatively small (i.e., none exceed 0.40). Furthermore, several of these differences (e.g., on Attention to Detail) may have little substantive meaning given they occurred at the upper end of the response scale. With regard to race/ethnicity, only one statistically significant difference was found. On average, Hispanics expected that Army work required more Innovation than White Non-Hispanics ($d = 0.50$).

Table 13.40. AWKS Scale Scores by Gender

Scale	d_{FM}	Male *M*	Male *SD*	Female *M*	Female *SD*
Dependability	0.23	4.25	0.86	4.45	0.67
Attention to Detail	**0.30**	4.22	0.87	4.48	0.69
Energy	**0.31**	4.14	0.98	4.45	0.67
Cultural Tolerance	**0.29**	4.15	0.93	4.42	0.73
Social Orientation	0.23	4.16	0.94	4.38	0.82
Self-Control	0.13	4.18	0.89	4.30	0.85
Adaptability/Flexibility	0.15	4.10	0.83	4.23	0.76
Cooperation	0.26	4.02	0.92	4.25	0.75
Stress Tolerance	**0.33**	3.99	0.99	4.31	0.93
Achievement/Effort	**0.34**	3.96	1.00	4.30	0.76
Persistence	0.12	4.02	0.90	4.13	1.00
Initiative	0.21	3.90	0.99	4.11	0.95
Leadership Orientation	-0.13	3.82	1.02	3.69	0.99
Concern for Others	**0.40**	3.43	1.10	3.87	0.84
Innovation	0.17	3.36	1.07	3.55	1.08
Independence	0.11	3.10	1.18	3.23	1.22

Note. $n_{Male} = 173$, $n_{Female} = 71$. d_{FM} = Effect size for Female-Male mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., Males) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

Table 13.41. AWKS Scale Scores by Race/Ethnic Group

Scale	d_{BW}	d_{HW}	White *M*	White *SD*	Black *M*	Black *SD*	White Non-Hispanic *M*	White Non-Hispanic *SD*	Hispanic *M*	Hispanic *SD*
Dependability	-0.04	0.29	4.28	0.82	4.24	0.89	4.25	0.85	4.50	0.62
Attention to Detail	0.20	-0.06	4.28	0.82	4.44	0.95	4.31	0.79	4.26	0.86
Energy	0.20	0.02	4.16	0.94	4.34	1.04	4.18	0.91	4.21	0.98
Cultural Tolerance	0.11	-0.01	4.19	0.89	4.29	0.96	4.21	0.88	4.21	0.88
Social Orientation	-0.09	0.08	4.20	0.91	4.12	1.08	4.22	0.91	4.29	0.87
Self-Control	0.16	-0.13	4.18	0.90	4.32	0.93	4.24	0.88	4.12	0.95
Adaptability/Flexibility	0.04	0.02	4.09	0.78	4.12	1.00	4.10	0.80	4.12	0.73
Cooperation	-0.09	-0.02	4.08	0.83	4.00	1.00	4.10	0.78	4.09	1.00
Stress Tolerance	0.12	-0.22	4.06	0.98	4.17	0.97	4.12	0.93	3.91	1.11
Achievement/Effort	0.01	0.19	4.01	0.99	4.02	1.04	4.02	0.98	4.21	0.95
Persistence	-0.06	0.13	4.05	0.88	4.00	1.05	4.07	0.88	4.18	0.83
Initiative	0.15	-0.20	3.96	0.94	4.10	1.07	3.99	0.94	3.79	1.01
Leadership Orientation	-0.06	0.12	3.79	0.94	3.73	1.32	3.79	0.95	3.91	1.00
Concern for Others	0.06	0.04	3.49	1.08	3.56	1.10	3.49	1.09	3.53	1.02
Innovation	0.15	**0.50**	3.36	1.05	3.51	1.16	3.28	1.03	3.79	1.04
Independence	0.16	0.22	3.13	1.15	3.32	1.31	3.10	1.12	3.35	1.35

Note. $n_{White} = 159$. $n_{Black} = 41$. $n_{White\ Non\text{-}Hispanic} = 136$. $n_{Hispanic} = 34$. d_{BW} = Effect size for Black-White mean difference. d_{HW} = Effect size for Hispanic-White Non-Hispanic mean difference. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* of referent group. Referent groups (e.g., White) are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

Scale Intercorrelations and Factor Structure

Table 13.42 shows intercorrelations among the AWKS scales. On average, the WVI scales showed moderate levels of intercorrelations (mean r = .46). Next, we used EFA to examine the factor structure underlying the 16 AWKS scales. Initially, we did not specify the number of factors to be extracted. Results indicated that three factors had eigenvalues over 1.0. The pattern matrix for this 3-factor solution revealed several cross loadings, and no interpretable factors. As such we conducted a series of follow-up EFAs that constrained the solution to three, two, and one factor, respectively. Each of the solutions from these analyses was dominated by a single common factor that reflects overly high expectations (based on the means presented earlier) about the temperament-related demands of Army work. The internal consistency reliability of a composite formed by averaging across the 16 AWKS scales was .88.

Table 13.42. Intercorrelations among AWKS Scores

Scale	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1 Dependability	-														
2 Attention to Detail	**.53**	-													
3 Energy	**.52**	**.56**	-												
4 Cultural Tolerance	**.56**	**.46**	**.49**	-											
5 Social Orientation	**.49**	**.54**	**.45**	**.53**	-										
6 Self-Control	**.44**	**.58**	**.41**	**.48**	**.45**	-									
7 Adaptability/Flexibility	**.49**	**.48**	**.45**	**.51**	**.40**	**.40**	-								
8 Cooperation	**.44**	**.51**	**.41**	**.50**	**.50**	**.45**	**.52**	-							
9 Stress Tolerance	**.44**	**.53**	**.49**	**.49**	**.46**	**.57**	**.42**	**.38**	-						
10 Achievement/Effort	**.41**	**.38**	**.37**	**.38**	**.39**	**.43**	**.27**	**.32**	**.45**	-					
11 Persistence	**.48**	**.56**	**.47**	**.55**	**.49**	**.48**	**.46**	**.39**	**.47**	**.39**	-				
12 Initiative	**.34**	**.42**	**.38**	**.45**	**.41**	**.48**	**.25**	**.22**	**.39**	**.37**	**.44**	-			
13 Leadership Orientation	**.28**	**.28**	**.22**	**.34**	**.30**	**.37**	**.36**	**.29**	**.31**	**.39**	**.34**	**.31**	-		
14 Concern for Others	**.24**	**.16**	**.29**	**.28**	**.28**	**.19**	**.30**	**.23**	**.13**	**.17**	**.15**	**.27**	**.17**	-	
15 Innovation	.11	**.18**	**.18**	**.18**	**.21**	.10	**.27**	.11	.06	**.24**	**.27**	**.21**	**.35**	**.38**	-
16 Independence	.05	.07	.09	.08	-.05	.04	**.17**	-.02	.06	.10	.10	**.15**	**.17**	**.32**	**.31**

Note. n = 245. Statistically significant correlations are bolded, p < .05 (two-tailed).

Relations between Expectations and Temperament

In the next set of analyses, we examined whether recruits' expectations regarding the temperament-related requirements of Army work can be differentiated from their temperament. Correlations between corresponding AWKS and WSI scale scores are presented in Table 13.43. Results revealed that recruits' expectations and preferences were generally unrelated (mean r = .06). Only two of the 16 scales showed significant correlations, and in these cases, the correlations were relatively small (r = .16 and .20). Thus, recruits' expectations regarding the temperament-related requirements of Army work appear to be distinct from their temperament.

Table 13.43. Correlations between Corresponding AWKS and WSI Scale Scores

Scale	*r*	Scale	*r*
Achievement/Effort	.08	Initiative	.04
Adaptability/Flexibility	.11	Innovation	.08
Attention to Detail	.11	Leadership Orientation	.02
Concern for Others	**.20**	Persistence	**.16**
Cooperation	-.04	Self-Control	.06
Dependability	.04	Social Orientation	.03
Energy	-.04	Stress Tolerance	-.08
Independence	.11	Cultural Tolerance	.11

Note. $n = 224$. Statistically significant correlations are bolded, $p < .05$ (one-tailed).

AWKS Fit Indices

Finally, we calculated Pearson correlations to assess profile similarity of recruits' AWKS scale scores to the WSSS scores.[84] The average correlation between recruits' AWKS profiles and the WSSS profile was .34 ($SD = .28$, range = -.34 to .92). The distribution of correlations across recruits was normally distributed. No significant differences were found on profile correlations for gender (Males: $M = 0.35$, $SD = 0.28$; Females: $M = 0.36$, $SD = 0.29$) or race/ethnicity (Whites: $M = 0.35$, $SD = 0.29$; Blacks: $M = 0.35$, $SD = 0.26$; Hispanics: $M = 0.29$, $SD = 0.32$).[85]

Next Steps for the AWKS

Because the Soldiers in the concurrent validation effort will have been in service for 18 to 36 months, it would not be appropriate to ask them to complete the AWKS based on what their expectations about the Army were prior to when they entered. As such, the AWKS will not be administered during the concurrent validation data collections. Nevertheless, we will include AWKS data gathered during the faking research and predictor field data collections in the Select21 attrition database to examine relationships between the AWKS (e.g., scale scores, composite, fit index) and attrition.

[84] Caution should be taken in interpreting these results because the AWKS and WSSS are scored on different metrics. Specifically, the AWKS scores are based on 5-point Likert-type ratings, whereas WSSS scale scores are on a 16-point metric based on transformed ranks (WSSS scale scores were calculated by subtracting each dimension's rank from 17) of mean SME ratings. Given these differences, no D^2 fit index was calculated. Thus, by examining Pearson r only, we are limiting assessment of profile similarity to differences in shape.

[85] Sample sizes for subgroup analyses were as follows: $n_{\text{Males}} = 163$; $n_{\text{Females}} = 65$; $n_{\text{Whites}} = 151$; $n_{\text{Blacks}} = 34$; $n_{\text{White Non-Hispanics}} = 129$; $n_{\text{Hispanics}} = 32$.

CHAPTER 14: CROSS INSTRUMENT ANALYSES

Chad H. Van Iddekinge, Christopher E. Sager, and Huy Le
HumRRO

Overview

In this chapter, we examine empirical relations among the Select21 criterion and predictor measures. There are two main reasons for performing cross instrument analyses. The first is that such analyses may identify criterion and/or predictor measures that are so highly related that it would not be a good use of time and resources to include both in the concurrent validation. Second, examining interrelations among criteria and predictors can provide construct-related validity evidence beyond that reported in the instrument-specific chapters. This is important because we want to provide support for using the criteria we developed as valid measures of current and expected future job performance (or work attitudes, in the case of the person-environment fit criteria). Likewise, we want to provide evidence for using the predictors we developed as measures of the critical knowledge, skills, and attributes (KSAs) identified by the Select21 job analysis. We begin by describing correlations among the criteria and then discuss the predictor intercorrelations.

Relations Among Criteria

We first describe the development and evaluation of a performance model of the Select21 criteria. This is followed by a discussion of relations between the performance model variables and three additional sets of criterion measures: (a) the MOS-specific criteria, (b) the attitudinal criteria, and (c) the future-oriented criteria.

Performance Model

Model Development

Our first step in analyzing the Select21 criterion data was to identify an a priori performance model to serve as an organizing framework. Given the similar aims of Select21 and Project A, we began by consulting the performance modeling results from the Project A research. After reviewing the various performance models evaluated in Project A, we concluded that the "leadership model" from the second-tour longitudinal sample (see J. P. Campbell & Knapp, 2001, p. 327) best represented the Select21 criterion space. This model comprises six performance components: (a) Core Technical Proficiency, (b) General Soldiering Proficiency, (c) Achievement and Effort, (d) Personal Discipline, (e) Physical Fitness/Military Bearing, and (f) Leadership. We did, however, modify the Project A model in a couple of ways. First, because we did not want to include MOS-specific criteria in the model (e.g., due to small sample sizes), we combined Core Technical Proficiency and General Soldiering Proficiency into a single component called General Technical Proficiency. Second, because the focus of Select21 is on developing measures to predict performance of first-term Soldiers who do not have permanent leadership responsibilities (whereas this particular Project A model focused on junior NCO performance), we renamed the Leadership component Teamwork. Thus, the final performance model comprised five components.

After identifying an a priori Select21 performance model, we then determined the criterion measures associated with each model component, again using the Project A results as a guide. We began by reviewing all of the non-ratings criteria and identified eight measures for inclusion in the model. These measures were the Army-wide Job Knowledge Test scores and Personnel File Form (PFF) Weapons Qualification scores (General Technical Proficiency component); PFF Awards, PFF Military Education, and Promotion Rate scores[86] (Achievement and Effort component); PFF Deviance (Personal Discipline component); Army Physical Fitness Test (APFT) scores (Physical Fitness and Military Bearing component); and Criterion Situational Judgment Test (CSJT) scores (Teamwork component).[87]

The next step was to determine how to best incorporate the Army-wide Current Observed Performance Rating Scales (COPRS) and composites into the performance model. To help guide our decision making, we first computed correlations between COPRS and composite scores and the eight non-ratings criteria. These correlations are displayed in Table 14.1. Two points are noteworthy about the correlations presented in this table. First, they are based on estimated COPRS ratings averaged across one supervisor and three peers. A description of the procedure used to estimate these correlations is provided in Appendix K. Second, because we were unable to estimate the reliability of many of the criterion measures (e.g., the self-report PFF scores), only observed correlations among criteria are reported in this and subsequent tables. As such, the reported correlations likely underestimate the magnitude of the "true" relationships among the underlying constructs. This is particularly relevant for the interpretation of relations between the COPRS and other criteria given the relatively low interrater reliability estimates of scores for some of the individual COPRS (see Chapter 3).

Our initial plan was to use scores from the ratings composites developed from confirmatory factor analysis of the COPRS data discussed in Chapter 3. However, we found that some individual rating scales within certain composites tended to be more related (theoretically and/or empirically) to the performance model component of interest than did other scales (see Table 14.1). For example, within the COPRS Physical Fitness and Self Development composite, Physical Fitness scores seemed more relevant to the Physical Fitness and Military Bearing component than did Personal and Professional Development scale scores. In fact, scores on the former scale demonstrated notably higher correlations with APFT scores ($r = .53$) than did scores on the latter scale ($r = .22$). The same was true for scales comprising the COPRS Technical Proficiency and Problem Solving composite. For instance, Common Task Performance and MOS-Specific Task Performance scale scores were more relevant (both theoretically and empirically) to other variables in the General Technical Proficiency model component (e.g., Job Knowledge Test scores) than were scores on COPRS Information Management and Problem Solving and Decision Making. As a result, we chose to use COPRS scores rather than the broader composite scores to round out the performance model.

---

[86] Promotion Rate scores were based on pay grade without consideration of MOS.

[87] CSJT scores are based on data from both final forms of the CSJT (see Chapter 5 for details). Specifically, scores from the final version of each form (13 items from Form A and 14 items from Form B) were transformed to *z*-scores and treated as equivalent forms. We note, however, that there were differences between the two forms (e.g., different patterns of relations with other criteria), and that these differences will likely attenuate correlations between the *z*-score based scores and other variables. Thus, the reported correlations are likely to be conservative estimates of the relationship between the CSJT and other criterion measures.

Table 14.1. Correlations between Army-Wide COPRS and Composite Scores and Scores from the Remaining Performance Model Variables

	Performance Model Variable							
COPRS/Composite Score	AW JK Test Scores	PFF Weapons Qualification	PFF Awards	PFF Deviance	APFT Scores	CSJT Scores	PFF Military Education	Promotion Rate Scores
Scales								
Common Task Performance	.23	.23	.17	-.11	.20	.10	.15	.16
MOS-Specific Task Performance	.17	.06	.06	-.08	.07	.09	.21	.18
Communication	.11	.24	.05	.07	.14	.18	.19	.10
Information Management	.05	.10	.02	-.05	.12	.09	.16	.09
Problem Solving and Decision Making	.02	.11	.11	-.11	.09	.13	.16	.18
Adaptation to Changes	.20	.10	.14	-.14	.15	.09	.16	.10
Effort and Initiative	.02	.01	.01	-.16	.13	.28	.17	.29
Professionalism and Personal Discipline	.03	-.05	-.03	-.17	.16	.21	.10	.23
Supports Peers	-.06	-.06	.01	-.06	.07	.22	.03	.06
Exhibits Tolerance	-.02	-.09	-.11	-.14	-.03	.27	-.01	.14
Personal and Professional Development	.13	.05	.00	-.21	.22	.18	.14	.22
Physical Fitness	.03	.04	.00	-.10	.53	.08	.03	.22
Mean Scale Scores	.10	.08	.05	-.16	.24	.23	.18	.25
Overall Effectiveness	.11	.07	.06	-.16	.24	.13	.15	.29
Composites								
Technical Proficiency and Problem Solving	.17	.18	.13	-.09	.16	.14	.22	.18
Effort and Initiative	.05	.02	-.01	-.19	.35	.23	.13	.31
Teamwork	-.04	-.09	-.06	-.12	.02	.29	.01	.13
Effort and Teamwork	-.01	-.06	-.03	-.17	.11	.32	.10	.25
Physical Fitness and Self Development	.08	.06	.00	-.17	.46	.13	.09	.26

Note. AW JK Test = Army-Wide Job Knowledge Test. PFF = Personnel File Form. CSJT = Criterion Situational Judgment Test. The procedure used to estimate correlations between the one-supervisor-three-peer COPRS ratings and the other criterion measures prevented us from estimating the standard errors of those correlations. As a result, the correlations in this table could not be tested for statistical significance.

Eight COPRS were selected for inclusion in the performance model. The Common Task Performance, MOS-Specific Task Performance, and Adaptation to Change scales were included as indicators of General Technical Proficiency component. The Effort and Initiative and Overall Effectiveness scales were deemed relevant to the Achievement and Effort component. The Physical Fitness scale was added to the Physical Fitness and Military Bearing component. And the Supports Peers and Exhibits Tolerance scales were included as indicators of the Teamwork component. Thus, the final performance model comprised 16 criterion measures—8 COPRS and 8 non-ratings variables. The variables included in each model component are discussed below.

Correlations Among Model Variables

Table 14.2 presents estimated correlations among the performance model variables. Also included at the bottom of the table are estimated correlations between the model variables and the COPRS we believed would not have a strong theoretical relationship with any of the model components. It is important to note that we did not account for the potential effects of method or rater variance on relations among the model variables. Method variance is a particular concern for estimating relations between the selected COPRS and the remaining variables given the common behavioral rating scales on which these scores are based. Rater variance is perhaps a particularly significant issue with the COPRS considering the amount of halo observed in these ratings (see Chapter 3 for details).

Beginning with the General Technical Proficiency component, scores on the three COPRS were only modestly related to the other criterion measures representing this component—Army-wide Job Knowledge Test scores and PFF Weapons Qualification scores. However, of the COPRS, these three scales generally had the strongest relations with the other two criteria in this model component. The smallest correlations among these variables were those between Weapons Qualification scores and MOS-Specific Task Performance and Adaptation scale scores (r = .06 and .10, respectively). It is also noteworthy that the three COPRS were as or more related to certain variables from other components than to variables within the General Technical Proficiency component. For example, Common Task Performance scores correlated .20 with APFT scores, and MOS-Specific Task Performance scores correlated .21 with PFF Military Education.

As for the Achievement and Effort component, scores on COPRS Effort and Initiative and Overall Effectiveness demonstrated decent correlations with Promotion Rate scores and PFF Military Education but were largely unrelated to PFF Awards (r = .01 and .06, respectively). In addition, there was actually a small negative correlation (r = -.07) between Promotion Rate scores and PFF Awards. Furthermore, the two COPRS were related to non-ratings measures within other model components. The most notable correlations were .28 between Effort and Initiative scale scores and CSJT scores and .24 between Overall Effectiveness scale scores and APFT scores.

The Personal Discipline component had only one indicator—PFF Deviance. As expected, Deviance correlated negatively with most of the other model variables. Deviance scores were most related to several variables from the Achievement and Effort component, including -.34 with Promotion Rate scores and -.16 with both COPRS Effort and Initiative and Overall Effectiveness scores. This suggests that Deviance might fit better in the Achievement and Effort component than as its own component.

Table 14.2. Correlations among Scores from the Performance Model Variables

Model Component/Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
General Technical Proficiency																
1. Army-Wide Job Knowledge Test																
2. PFF Weapons Qualification	.26															
3. COPRS Common Task Perf scale	.23	.23														
4. COPRS MOS-Specific Task Perf scale	.17	.06	.59													
5. COPRS Adaptation to Changes scale	.20	.10	.51	.52												
Achievement and Effort																
6. PFF Awards	.14	.17	.17	.06	.14											
7. Promotion Rate	.07	.05	.16	.18	.10	-.07										
8. PFF Military Education	.17	.06	.15	.21	.16	.25	.09									
9. COPRS Effort and Initiative scale	.02	.01	.53	.56	.49	.01	.29	.17								
10. COPRS Overall Effectiveness scale	.11	.07	.63	.66	.61	.06	.29	.15	.67							
Personal Discipline																
11. PFF Deviance	.04	.04	-.11	-.08	-.14	.02	-.34	.01	-.16	-.16						
Physical Fitness/Military Bearing																
12. APFT score	.02	.19	.20	.07	.15	.12	.11	.10	.13	.24	-.09					
13. COPRS Physical Fitness scale	.03	.04	.38	.24	.30	.00	.22	.03	.38	.52	-.10	.53				
Teamwork																
14. CSJT score	-.05	-.16	.10	.09	.09	-.15	.08	.12	.28	.13	-.09	-.10	.08			
15. COPRS Supports Peers scale	-.06	-.06	.31	.34	.33	.01	.06	.03	.45	.49	-.06	.07	.23	.22		
16. COPRS Exhibits Tolerance scale	-.02	-.09	.23	.19	.24	-.11	.14	-.01	.38	.33	-.14	-.03	.14	.27	.47	
Excluded COPRS																
Communication	.11	.24	.52	.41	.37	.05	.10	.19	.43	.52	.07	.14	.34	.18	.27	.25
Information Management	.05	.10	.51	.55	.52	.02	.09	.16	.49	.61	-.05	.12	.27	.09	.32	.23
Problem Solving & Decision Making	.02	.11	.56	.60	.48	.11	.18	.16	.52	.69	-.11	.09	.31	.13	.36	.23
Professionalism and Personal Discipline	.03	-.05	.42	.49	.42	-.03	.23	.10	.64	.58	-.17	.16	.41	.21	.45	.45
Personal and Professional Development	.13	.05	.55	.49	.48	.00	.22	.14	.56	.67	-.21	.22	.44	.18	.31	.33

Note. PFF = Personnel File Form. CSJT = Criterion Situational Judgment Test. Excluded COPRS were those we felt did not have a strong theoretical relationship with any of the performance model components. The procedure to estimate correlations between the one-supervisor-three-peer COPRS ratings and the other criterion measures prevented us from estimating the standard errors of those correlations. As a result, the correlations in this table could not be tested for statistical significance.

The two criteria that comprised the Physical Fitness and Military Bearing component—APFT scores and COPRS Physical Fitness scores—were highly correlated ($r = .53$). In fact, this was the largest correlation between the COPRS and the eight non-ratings criteria in the model. It is also noteworthy that Physical Fitness scores correlated higher with APFT scores than with the other COPRS scores, which provides some evidence for the discriminant validity of this COPRS dimension.

Finally, the Teamwork component included CSJT scores and COPRS Supports Peers and Exhibits Tolerance. As discussed, although CSJT scores correlated with scores on these two COPRS, performance on the CSJT was slightly more related to COPRS Effort and Initiative scores. The somewhat larger correlations between the CSJT and these three COPRS make sense given that many of the CSJT items were designed to assess judgment in relation to job dimensions relevant to these rating scales—namely, relating to peers, teamwork, self-management, and self-directed learning (see Chapter 5). It is also interesting that CSJT scores were positively related to all COPRS scores but negatively related to the remaining model variables (e.g., $r = -.16$ with PFF Weapons Qualification). Again, these relationships should be interpreted somewhat cautiously given the limitations of the z-score based CSJT scores used in these analyses.

Modeling Analyses

Our original intent was to use a covariation matrix of the variables in Table 14.2 as input for a confirmatory factor analysis to assess the fit of the performance model. We subsequently decided, however, not to conduct further performance model analyses using the field test data. The main reason for this decision was that the pairwise sample sizes for the 16 performance model variables varied greatly ($n = 142$ to 313), and only 91 Soldiers had data on all variables. As such, we felt that this sample was insufficient for modeling, particularly because we would have had to incorporate additional parameters to model both method and rater factors. Two additional factors also played a role in this decision. First, one of the primary goals of the cross instrument analyses was to determine whether there is redundancy among the criterion measures. Examination of the intercorrelations indicated that this was not the case, and the modeling results would be unlikely to change this conclusion. In other words, the results of the performance modeling would not lead us to recommend excluding a criterion measure from the concurrent validation. Second, another goal of the cross instrument analyses was to further assess the construct-related validity of the Select21 criterion measures. Although evaluating a performance model would provide a more complete assessment of construct validity, examination of intercorrelations among the criteria allowed us to draw some conclusions about the validity of these measures.

Nevertheless, we do plan to assess the fit of the identified performance model during the concurrent validation, as we anticipate having a larger sample size on which to conduct such analyses. Modeling relations among the concurrent validation criteria will be important for several reasons. For example, as mentioned above, modeling analyses can provide a more complete assessment of the construct-related validity of the performance measures. Establishing the construct validity of the criteria is particularly important for evaluating the criterion-related validity of the predictors we are developing for this project. Furthermore, the modeling analyses may provide support for the existence of a more parsimonious set of criterion variables (e.g., the five factors from the a priori performance model). Once identified, these variables could be used to assess the validity of the predictor measures.

Relations between Performance Model Variables and Other Criteria

MOS-Specific Criteria

We now briefly discuss relations between the performance model variables and three sets of additional Select21 criterion measures. Table 14.3 displays correlations between the model variables and the MOS-specific criteria. Several findings are noteworthy. For one, Army-wide Job Knowledge Test scores correlated highly with 11B and 31U Job Knowledge Test scores (r = .71 and .59, respectively). It is also interesting to note some of the differential relations between scores on the two job knowledge tests and the other criteria. For instance, the Physical Fitness and Military Bearing component variables were positively related to 11B Knowledge Test scores but negatively related to 31U Knowledge Test scores. These results suggest that the two knowledge tests are measuring somewhat different constructs.

Table 14.3. Correlations between Scores from the Performance Model Variables and Scores from the MOS-Specific Criteria

	MOS-Specific Criteria			
Model Component/Variable	11B Job Knowledge Test Scores	31U Job Knowledge Test Scores	11B COPRS Composite Scores	11B FX Composite Scores
General Technical Proficiency				
Army-Wide Job Knowledge Test	.71	.59	.10	-.10
PFF Weapons Qualification	.18	.32	.21	-.06
COPRS Common Task Perf scale	.15	.16	.60	.37
COPRS MOS-Specific Task Perf scale	.14	.07	.72	.37
COPRS Adaptation to Changes scale	.01	.27	.46	.27
Achievement and Effort				
PFF Awards	.19	.35	.23	.37
Promotion Rate	.17	-.05	.17	.25
PFF Military Education	.17	.21	.39	.08
COPRS Effort and Initiative scale	.03	.14	.55	.44
COPRS Overall Effectiveness scale	.12	.19	.71	.46
Personal Discipline				
PFF Deviance	-.15	-.03	.10	.13
Physical Fitness/Military Bearing				
APFT score	.20	-.19	.15	.11
COPRS Physical Fitness scale	.07	-.39	.40	.17
Teamwork				
CSJT score	.25	.09	.08	-.17
COPRS Supports Peers scale	-.13	.26	.37	.31
COPRS Exhibits Tolerance scale	-.06	.09	.36	.18

Note. MOS-specific criteria not shown in this table (e.g., 31U ratings) were excluded due to small sample sizes. The procedure to estimate correlations between the one-supervisor-three-peer COPRS ratings and the other criterion measures prevented us from estimating the standard errors of those correlations. As a result, the correlations in this table could not be tested for statistical significance.

As for the MOS-specific performance ratings, it is not surprising the 11B COPRS composite scores, and to a lesser extent the 11B Expected Future Performance (FX) Rating Scales composite, were highly related to the Army-wide COPRS (e.g., $r = .71$ and .46, respectively, with COPRS Overall Effectiveness scores). The two MOS-specific ratings composites were also related to several of the non-ratings criteria, such as PFF Awards and Promotion Rate scores.

Attitudinal Criteria

The Select21 criterion space also includes measures of several attitudinal variables (assessed by the Army Life Survey [ALS]) that the person-environment fit measures are expected to predict (see Chapter 7 for details). Correlations between the performance model variables and ALS attitudinal scale scores are shown in Table 14.4. Overall, the performance model criteria were only modestly related to the attitudinal criteria, with relatively few correlations exceeding .20. Interestingly, CSJT scores demonstrated the strongest and most consistent relations with the attitudinal criteria. For example, the CSJT correlated .40 with Army Fit scale scores, -.34 with Attrition Cognition scale scores, and between .24 and .34 with all of the Satisfaction with the Army scale scores. One possible interpretation of the CSJT-ALS relations is that Soldiers who are more satisfied in the Army took the field test more seriously and devoted more thought to their CSJT responses.

As for the other performance model variables, the COPRS tended to demonstrate small, positive correlations with the attitudinal criteria, with Exhibits Tolerance scale scores being most related to the ALS measures (e.g., $r = .23$ with Satisfaction with Peers scores). Conversely, several of the PFF criterion variables, such as Weapons Qualification and Award scores, were negatively related to attitude scores. Army-wide Job Knowledge Test scores also tended to correlate negatively with the ALS measures, including -.29 with Satisfaction with Work Itself scores and -.22 with Satisfaction with Pay and Benefits scores.

Future-Oriented Criteria

The final relations of interest were those between the performance model variables and the future-oriented criterion measures, which include the Army-wide FX composite and the three composites from the Future Army Life Survey (FALS). These correlations are presented in Table 14.5. As discussed in Chapter 3, there were moderate to large correlations between COPRS scores and FX composite scores. FX composite scores were also moderately related to several non-ratings criteria, including Promotion Rate scores ($r = .22$), PFF Military Education ($r = .24$), and APFT scores ($r = .23$). In general, the FALS composite scores did not correlate very highly with the performance model variables. The one notable exception was that Future Fit scores were modestly correlated with some of the model variables, including .25 with CSJT scores, .22 with Army-wide Job Knowledge Test scores, and .20 with PFF Military Education.

Table 14.4. Correlations between Scores from the Performance Model Variables and Scores from the ALS

	Satisfaction with Army						Organizational Commitment			Perceived Fit			Army Career Intentions			
Model Component/Variable	Super	Peers	Work Itself	Prom	Pay/ Benefits	Army General	Affect	Cont	Norm	MOS Fit	Army Fit	Percvd Stress	Attrit Cogn	Reenlist Intent	Career Intent	Army Values
General Technical Proficiency																
Army-Wide Job Knowledge Test	-.10	-.05	-.29	-.18	-.22	-.07	.00	-.05	-.10	-.10	.01	-.02	-.16	-.04	.01	.04
PFF Weapons Qualification	-.19	-.04	-.14	-.16	-.24	-.10	-.03	-.07	-.05	-.10	-.05	.07	.06	-.05	-.09	-.05
COPRS Common Task Perf scale	.06	.01	.10	.03	-.02	-.04	-.14	-.03	.07	-.01	-.02	-.04	.12	-.12	-.08	.11
COPRS MOS-Specific Task Perf scale	.08	.03	.01	.03	-.11	-.09	-.06	-.07	-.07	.01	.01	.00	-.05	-.01	.00	-.11
COPRS Adaptation to Changes scale	.15	.16	.08	.01	-.17	.02	.06	-.05	.04	.09	.13	-.12	-.17	.09	.13	.05
Achievement and Effort																
PFF Awards	-.17	-.02	-.05	-.16	-.19	-.10	-.01	-.01	-.08	.00	-.09	.08	-.18	-.04	-.02	.06
Promotion Rate	.13	.02	.10	.20	.07	.04	.03	-.05	-.02	.07	.09	-.05	-.05	.08	.05	.03
PFF Military Education	.03	.04	.04	.05	.06	.12	.11	-.04	-.05	.13	.14	-.14	-.12	.07	.06	.08
COPRS Effort and Initiative scale	.14	.13	.16	.15	.02	.17	.15	.10	.08	.12	.21	-.18	-.11	.15	.14	.11
COPRS Overall Effectiveness scale	.17	.13	.06	.11	-.01	.05	.10	-.07	.02	.08	.15	-.18	-.12	.05	.07	.04
Personal Discipline																
PFF Deviance	-.07	-.04	-.05	-.02	.00	.03	-.02	.12	.09	.04	.01	-.05	.05	.06	.07	.02
Physical Fitness/Military Bearing																
APFT score	-.02	-.03	.01	.03	-.11	.02	.10	-.04	.03	-.01	.06	-.08	-.01	.03	.04	.08
COPRS Physical Fitness scale	.03	.06	.02	.08	.02	.12	.13	.04	.08	.03	.15	-.16	-.09	.10	.05	.06
Teamwork																
CSJT score	.31	.34	.30	.24	.25	.31	.32	.08	.06	.18	.40	-.23	-.34	.19	.16	.35
COPRS Supports Peers scale	.19	.17	.09	.03	.12	.10	.08	-.03	.02	.03	.14	-.06	.01	.07	.03	.08
COPRS Exhibits Tolerance scale	.23	.23	.14	.21	.19	.17	.18	-.01	.03	.09	.18	-.08	-.07	.07	.04	.11

Note. Super = Satisfaction with Supervision. Prom = Satisfaction with Promotions. Affect = Affective Commitment. Cont = Continuance Commitment. Norm = Normative Commitment. Attrit Cogn = Attrition Cognitions. The procedure to estimate correlations between the one-supervisor-three-peer COPRS ratings and the other criterion measures prevented us from estimating the standard errors of those correlations. As a result, the correlations in this table could not be tested for statistical significance.

Table 14.5. Correlations between Scores from the Performance Model Variables and Scores from the Future-Oriented Criteria

Model Component/Variable	FX Rating Composite	Future Army Life Survey Composite: Future Fit	Future Stress	Future Continue
General Technical Proficiency				
Army-Wide Job Knowledge Test	.14	.22	-.01	.06
PFF Weapons Qualification	.10	.06	-.08	.07
COPRS Common Task Perf scale	.56	.07	-.01	.10
COPRS MOS-Specific Task Perf scale	.57	.05	.04	-.07
COPRS Adaptation to Changes scale	.52	.11	-.08	.01
Achievement and Effort				
PFF Awards	.10	.00	-.02	.02
Promotion Rate	.22	.06	.03	-.13
PFF Military Education	.24	.20	.01	.06
COPRS Effort and Initiative scale	.55	.04	.05	-.02
COPRS Overall Effectiveness scale	.71	.12	-.01	-.05
Personal Discipline				
PFF Deviance	-.10	.07	-.12	.15
Physical Fitness/Military Bearing				
APFT score	.23	.04	-.03	-.03
COPRS Physical Fitness scale	.39	.08	-.06	-.03
Teamwork				
CSJT score	.05	.25	-.01	.13
COPRS Supports Peers scale	.29	.02	.01	-.03
COPRS Exhibits Tolerance scale	.26	.07	-.04	-.02

Note. Future Continue = Future Continuance. The procedure used to estimate correlations between the one-supervisor-three-peer COPRS ratings and the other criterion measures prevented us from estimating the standard errors of those correlations. As a result, the correlations in this table could not be tested for statistical significance.

Summary

Examination of relations among the Select21 criterion measures revealed several informative results. One of the main conclusions is that, with the notable exception of the COPRS and composites, there appears to be minimal overlap among criterion scores. The cross instrument analysis results also provide some evidence for the construct-related validity of the criterion measures. Perhaps the most salient example is the strong correspondence between APFT scores and COPRS Physical Fitness scores. It is also encouraging that CSJT scores were slightly more related to the COPRS performance dimensions similar to those targeted by the CSJT items.

Nonetheless, correlations among criteria thought to tap the same or similar aspects of the performance domain tended to be rather modest. In fact, there were several instances in which criteria within a given performance model component demonstrated little or no relationship and/or correlated more highly with one or more criteria from another model component. Results also suggested that the five-component performance model may not be the most appropriate

model for the Select21 criteria. For example, perhaps the Achievement and Effort and Personal Discipline components should be combined given the relatively strong relations between PFF Deviance and several variables within the Achievement and Effort component.

It will also be interesting to see how CSJT scores correlate with other variables when a single test form is used during the concurrent validation. As discussed, CSJT scores were both theoretically and empirically related to both the Achievement and Effort and Teamwork components. The consistent relations between performance on the CSJT and scores on the attitudinal criteria should also be attended to during the validation research. Finally, one of the main performance modeling challenges during the concurrent effort will likely be how to best account for the method factors (and rater factors in the case of the COPRS) that might account for some of the observed covariation among certain criteria.

Relations Among Predictors

We now examine relations among the predictor measures being developed for this project. Results are discussed by the following predictor clusters: (a) cognitive ability, psychomotor ability, and judgment; (b) education, training, and experience; (c) temperament; (d) person-environment (P-E) fit needs measures; (e) expectations about the Army; and (f) P-E fit index scores. These predictors yield a very large number of scale and composite scores. Thus, to facilitate interpretation, we provide two tables for each predictor cluster. The first set of tables displays correlations among all scale and composite scores from measures within a given cluster (e.g., all temperament scales and composites). The second set of tables display correlations between scales and composites within a predictor cluster and the scales/composites of predictors from most or all other clusters, whereby the names of the measures from a given cluster appear in the columns and names of the remaining predictors appear in the rows.

We note two additional characteristics of correlations presented in this section. First, all analyses were performed using pairwise deletion. As such, the sample sizes on which the correlations are based vary considerably within some of the tables. Second, as with the criterion measures, we are unable to estimate the reliability of many of the predictors, and thus we report only observed correlations among predictor scores. Therefore, the reported correlations likely underestimate the magnitude of the "true" relationships among the underlying predictor constructs.

Cognitive Ability, Psychomotor Ability, and Judgment

The first set of predictors includes measures of cognitive ability, psychomotor ability, and judgment. The cognitive ability scores consist of five Armed Services Vocational Aptitude Battery (ASVAB) composites. Each composite consists of a subset of ASVAB tests shown in Table 14.6. The Armed Forces Qualification Test (AFQT) composite is used operationally for Soldier selection and is viewed as a reasonable measure of general cognitive aptitude. The remaining composites have been used in previous research investigating the construct-related validity of predictors designed to supplement the ASVAB (e.g., J. P. Campbell & Knapp, 2001). The correlations among these composites are consistent with the overlap in tests composing them and their theoretical relationships (see Table 14.7).

Table 14.6. ASVAB Tests Comprising each Cognitive Ability Composite

Composite	Constituent ASVAB Test(s)
AFQT	Arithmetic Reasoning, Math Knowledge, Word Knowledge, Paragraph Comprehension
Verbal	Word Knowledge, Paragraph Comprehension, General Science
Quantitative	Arithmetic Reasoning, Math Knowledge
Technical	Auto Information, Shop Information, Mechanical Comprehension, Electronics Information
Spatial	Assembling Objects

The two psychomotor tests—Target Shoot and Target Tracking—and the Predictor Situational Judgment Test (PSJT) are experimental measures developed for this project. The psychomotor tests yield two scale scores: one that measures accuracy (Psychomotor Precision composite) and one that measures reaction time (Time-to-Fire score).[88] While lower scores represent superior performance for both psychomotor measures, we recoded their scores such that positive correlations with other instruments represent positive covariation relative to performance. The PSJT yields an Overall Judgment score for each of the two forms of the test. The Judgment scores used in the cross instrument analyses were based on data from both final forms of the PSJT (see Chapter 10 for details). As with the CSJT, scores from the final version of each PSJT form (12 items from Form A and 14 items from Form B) were transformed to z-scores and treated as equivalent forms. We note, however, that like the CSJT, there were differences between the two PSJT forms (e.g., different patterns of relations with other predictors), and that these differences will likely attenuate correlations between Overall Judgment scores and other variables. Thus, the reported correlations are likely to be conservative estimates of the relationship between the PSJT and other predictors.

Table 14.7 displays intercorrelations among scores from the abilities and judgment measures. The two psychomotor scale scores were consistently related to cognitive ability scores such that recruits with higher cognitive ability tended to have higher psychomotor ability scores. In fact, psychomotor scores correlated almost the same with the ASVAB Technical composite (r = .36 and .37) as they did with one another (r = .35). As for the PSJT, Overall Judgment scores correlated around .20 with all of the cognitive ability measures (except the ASVAB Technical composite), but were unrelated to the psychomotor scores. Thus, verbal and quantitative ability appear to be most relevant for performance on the PSJT, whereas the more "practical" abilities measured by the Technical composite appear to be most relevant to psychomotor ability.

[88] Given the male-female score differences discussed in Chapter 11, we examined correlations between the psychomotor scores and other predictors using data from male recruits only and data from the entire sample, males and females. The two sets of results were very similar. For example, the average (absolute) correlation between the two psychomotor scores and the remaining predictors were almost identical for the full and male only samples. In addition, although there were some differences in the pattern of relations between the psychomotor scales and other predictors in the two samples, the two sets of correlations (i.e., between psychomotor scores and the other predictors in the full and males only samples) were highly related (e.g., r = .76 for Time-to-Fire scale scores). Based on these observations, we chose to report results using psychomotor data from the entire available sample.

Table 14.7. Correlations among Cognitive Ability, Psychomotor Ability, and Judgment Predictor Measures

Predictor	1	2	3	4	5	6	7	8
ASVAB Composite Scores								
1. AFQT	—							
2. Verbal	**.79**	—						
3. Quantitative	**.84**	**.43**	—					
4. Technical	**.50**	**.60**	**.33**	—				
5. Spatial	**.37**	**.26**	**.44**	**.37**	—			
Psychomotor Ability Test Scores								
6. Time-to-Fire Scale	**.26**	**.28**	**.22**	**.36**	**.25**	—		
7. Precision Composite	**.26**	**.29**	**.18**	**.37**	**.21**	**.35**	—	
8. PSJT Overall Judgment Score	**.25**	**.21**	**.20**	.06	**.17**	.04	.04	—

Note. n = 455 to 671. PSJT = Predictor Situational Judgment Test. Bolded correlations are significant, $p < .05$ (two-tailed).

Correlations between scores on this set of predictors and scores on the remaining Select21 predictors are shown in Table 14.8. The two psychomotor scale scores were, in general, relatively independent from the other predictors. Interestingly, psychomotor scores were most related to vocational interest scores from the Work Preferences Survey (WPS) and Interest Finder Questionnaire (IFQ). For example, Time-to-Fire and Precision Composite scores negatively correlated -.16 and -.17, respectively, with IFQ Social scores and -.17 and -.18 with IFQ Conventional scores. Additionally, psychomotor scores were positively related to IFQ Investigative scores (r = .09 and .17).

PSJT Overall Judgment scores correlated as or more strongly with several of the noncognitive predictors than with the cognitive predictors. Most notably, the PSJT was consistently related to scores on the Rational Biodata Inventory (RBI). The strongest relations were with RBI Hostility to Authority (r = -.39), Cultural Tolerance (r = .34), and Internal Locus of Control (r = .31) scale scores. These relations were not unexpected given that the PSJT was designed to measure KSAs conceptually related to certain temperament constructs. The negative correlation with Hostility to Authority is particularly sensible given that many of the PSJT items assess willingness to "do the right thing" in situations where there are fairly clear rules to guide one's behavior.

Correlations between PSJT Overall Judgment scores and scores on the other Select21 temperament instrument, the Work Suitability Inventory (WSI), were small and generally nonsignificant.[89] Note, however, that the WSI scale scores are substantially ipsative (see Chapter 9 for details). For example, positive correlations between WSI scale scores and other variables indicate that Soldiers who have high standing on a given scale tended to view themselves as more capable of performing a given type of work (linked to a particular trait) relative to other types of work. In other words, the magnitude of WSI scale scores does not necessarily reflect Soldiers' standing on a given temperament variable, but rather how well they feel they could perform work that requires a certain trait relative to other types of work associated with other traits. We also note that the ipsative format of the WSI does not allow us to estimate the internal consistency reliability of its scale scores. Thus, we do not know the extent to which this type of

[89] As discussed in Chapter 10, the PSJT was also scored to measure certain temperament variables. Correlations between the PSJT temperament scores and the remaining predictors are discussed later and can be found in Tables 14.11 through 14.13.

measurement error contributes to the modest WSI relationships. Taken together, correlations between the WSI and other predictor measures should be interpreted with caution.

Table 14.8. Correlations between Ability and Judgment Predictors and the Other Predictor Measures

	ASVAB Composite Scores					Psychomotor Test Scores		
Predictor	AFQT	Verbal	Quant	Technical	Spatial	Time-to-Fire	Precision Composite	PSJT
Education Tier	-.08	.02	**-.14**	**.13**	-.04	.07	.00	.01
REPETE Scale Scores								
Computer Courses Taken	**.12**	**.09**	**.16**	.02	.07	.03	.05	-.03
Mean Level of Mastery	**.20**	**.14**	**.22**	.04	.08	**.15**	**.12**	-.02
General Computer Skills	**.19**	**.12**	**.23**	.00	.06	**.09**	**.11**	.01
Basic Computer Certs.	**-.10**	**-.11**	**-.09**	-.07	**-.14**	**-.09**	.03	**-.10**
Advanced Computer Certs.	.02	-.01	.01	.04	-.06	-.05	.03	-.04
RBI Scale Scores								
Peer Leadership	**.16**	**.20**	**.09**	**.08**	.01	.04	.06	**.16**
Cognitive Flexibility	**.22**	**.25**	**.15**	**.10**	.06	.06	.08	**.22**
Achievement Orientation	**.09**	-.02	**.09**	-.08	-.01	-.08	-.03	**.24**
Fitness Motivation	**.10**	**.09**	.08	.13	.06	**.10**	**.10**	.06
Diplomacy	.03	.03	-.01	.01	.00	.00	-.01	**.21**
Stress Tolerance	**.15**	**.16**	**.13**	**.19**	.08	**.10**	**.10**	.06
Hostility to Authority	**-.15**	**-.12**	**-.17**	-.01	**-.12**	-.01	-.02	**-.39**
Self-Esteem	.07	.06	.07	.08	.02	-.01	.02	**.24**
Narcissism	.00	-.04	-.03	**-.09**	-.01	-.03	-.06	**.13**
Cultural Tolerance	**.10**	**.12**	.06	.04	**.12**	.07	.02	**.34**
Internal Locus of Control	**.19**	**.17**	**.14**	**.09**	.07	**.09**	.07	**.31**
Army Identification	.01	.08	-.05	**.11**	.05	.06	.07	**.19**
Respect for Authority	.07	.02	**.09**	-.07	-.01	-.02	-.02	**.10**
Lie Scale	**-.12**	**-.13**	-.07	-.04	.01	-.05	-.02	**-.11**
WSI Scale Scores								
Achievement/Effort	.00	-.08	.03	**-.08**	-.05	**-.08**	**-.09**	-.02
Adaptability/Flexibility	-.03	-.03	**-.08**	**-.08**	.00	.01	-.04	.00
Attention to Detail	.02	.00	.01	.05	-.02	-.03	-.02	.02
Concern for Others	**-.10**	**-.10**	**-.09**	**-.23**	**-.17**	-.05	**-.11**	-.01
Cooperation	**-.23**	**-.20**	**-.16**	**-.23**	**-.12**	**-.10**	**-.10**	**-.11**
Dependability	.08	.01	.07	-.01	**.11**	-.07	-.01	**.10**
Energy	.04	-.01	.00	.05	.01	.00	.01	-.04
Independence	.08	**.10**	.05	**.14**	.05	.03	.01	.02
Initiative	.00	.02	.00	.04	-.03	-.08	-.04	-.02
Innovation	.05	**.09**	.04	**.10**	.05	.02	-.02	.00
Leadership Orient	.05	.04	.07	.05	**.10**	.05	.04	.00
Persistence	.06	.04	**.09**	**.17**	**.10**	.07	.07	.02
Self-Control	**.12**	**.14**	**.08**	**.16**	.06	.05	**.11**	**.10**
Social Orientation	**-.12**	-.06	**-.10**	**-.08**	-.05	.04	.07	-.07
Stress Tolerance	.08	**.11**	.06	**.10**	.01	.06	**.12**	.01
Cultural Tolerance	-.09	-.06	-.05	**-.12**	-.02	.05	-.01	.02

Table 14.8. (continued)

	ASVAB Composite Scores					Psychomotor Test Scores		
Predictor	AFQT	Verbal	Quant	Technical	Spatial	Time-to-Fire	Precision Composite	PSJT
WVI Composite Scores								
Growth	.03	-.01	.03	-.03	.06	-.03	.01	**.14**
Status	-.03	-.05	-.07	-.03	.01	-.02	.00	.09
Stimulation	-.02	.00	-.05	.00	.04	-.00	.05	.01
Comfort	.02	.00	-.01	.00	-.05	.07	-.03	.01
Altruism	**-.10**	**-.10**	**-.11**	**-.12**	-.07	-.07	-.03	**.11**
Self-Direction	.03	.05	.00	.03	-.03	.00	.00	.01
WPS Scale Scores								
Realistic	**-.10**	-.03	**-.14**	**.27**	.06	.07	**.11**	-.03
Investigative	**.11**	**.09**	**.08**	.03	.00	.04	.07	**.20**
Artistic	-.05	-.01	-.06	-.03	.01	-.04	-.03	.05
Social	**-.11**	**-.12**	**-.10**	**-.25**	**-.11**	**-.13**	**-.13**	**.19**
Enterprising	-.01	-.05	-.03	**-.14**	-.08	-.05	-.04	**.16**
Conventional	**-.12**	**-.27**	-.02	**-.26**	-.05	**-.17**	-.14	**.13**
IFQ Scale Scores								
Realistic	-.02	-.01	-.04	**.30**	.07	-.01	**.09**	-.01
Investigative	**.16**	**.24**	**.11**	**.17**	.05	**.09**	**.17**	**.20**
Artistic	-.02	.04	-.05	**-.10**	-.03	**-.08**	-.04	**.13**
Social	-.06	**-.08**	**-.09**	**-.26**	**-.15**	**-.16**	**-.17**	**.23**
Enterprising	.06	-.02	.03	**-.10**	**-.09**	-.07	-.01	.05
Conventional	**-.11**	**-.28**	.00	**-.29**	-.03	**-.17**	**-.18**	**.10**

Note. n = 222 to 670. Quant = Quantitative composite. Bolded correlations are significant, $p < .05$ (two-tailed).

As with the WSI, correlations between PSJT Overall Judgment scores and composite scores from the Work Values Inventory (WVI) were quite modest. Conversely, there were several significant correlations between the PSJT and vocational interest scales. The most notable relations were between judgment scores and WPS and IFQ Social scale scores (r = .23 and .20, respectively) and Investigative scale scores (both rs = .20).

Education, Training, and Experience

The next group of predictors focuses on education, training, and experience. The variables include education tier, which is an operational measure used to select enlisted Soldiers,[90] and the REPETE, which is an experimental predictor that yields five scale scores (see Chapter 12 for a description). Intercorrelations among these measures are displayed in Table 14.9. In general, REPETE scale scores were negatively related to education tier. This is a sensible relationship because Tier I recruits are more educated than Tier II recruits (there were no Tier III recruits in this sample). Nonetheless, these correlations were quite modest (i.e., $r \geq -.15$), which suggests that the REPETE provides unique information about education, training, and experience beyond education tier.

[90] Education tier is coded such that lower values indicate higher education. Specifically, Tier 1 = high school diploma graduate, Tier 2 = alternative credential holder (e.g., HS equivalency), and Tier 3 = non-high school graduate and no alternative credential.

Table 14.9. Correlations among Education, Training, and Experience Predictor Measures

Predictor	1	2	3	4	5	6
1. Education Tier						
REPETE Scale Scores	—					
2. Computer Courses Taken	**-.15**	—				
3. Mean Level of Mastery Rating	-.06	**.31**	—			
4. General Computer Skills	**-.15**	**.52**	**.86**	—		
5. Basic Computer Certifications	.01	**.12**	**.27**	**.29**	—	
6. Advanced Computer Certifications	-.02	**.14**	**.29**	**.30**	**.58**	—

Note. n = 545 to 608. Bolded correlations are significant, $p < .05$ (two-tailed).

Table 14.10 shows correlations between scores from this set of predictors and scores from the remaining predictor measures. REPETE scores were significantly related to several predictor constructs. For example, Computer Courses, Level of Mastery, and General Computer Skills scores correlated between .15 and .25 with ASVAB AFQT, Verbal, and Quantitative composite scores. In contrast, Basic and Advanced Computer Certifications scores had no or a negative relationship to cognitive ability. As for the noncognitive measures, Level of Mastery and General Computer Skills correlated significantly with several of the RBI scales, including Peer Leadership, Cognitive Flexibility, Achievement, and Fitness Motivation. The Mastery-Peer Leadership correlation of .26 was the highest between the REPETE scales and the other predictors. These two REPETE scales were also consistently related to several of the vocational interest scales, particularly Investigative, Artistic, and Enterprising. The correlation with Investigative scores is not surprising given that individuals with such interests tend to seek out educational opportunities.

Table 14.10. Correlations between Education, Training, and Experience Predictors and the Other Predictor Measures

		REPETE Scale Scores				
Predictor	Education Tier	Computer Courses	Level of Mastery	General Computer Skills	Basic Computer Certs.	Advanced Computer Certs.
ASVAB Composite Scores						
AFQT	-.08	**.12**	**.20**	**.19**	**-.10**	.02
Verbal	.02	**.09**	**.14**	**.12**	**-.11**	-.01
Quantitative	**-.14**	**.16**	**.22**	**.23**	**-.09**	.01
Technical	**.13**	.02	.04	.00	-.07	.04
Spatial	-.04	.07	.08	.06	**-.14**	-.06
PSJT Judgment Score	.01	-.03	-.02	.01	**-.10**	-.04
Psychomotor Test Scores						
Time-to-Fire	.07	.03	**.15**	**.09**	**-.09**	-.05
Precision Composite	.00	.05	**.12**	**.11**	-.03	.03
RBI Scale Scores						
Peer Leadership	.08	.05	**.26**	**.24**	.02	.02
Cognitive Flexibility	-.01	.07	**.25**	**.22**	.00	.05
Achievement Orientation	**-.08**	.04	**.15**	**.18**	.01	.01
Fitness Motivation	.01	.08	**.15**	**.13**	.01	.03
Diplomacy	**.10**	-.01	.05	.07	-.01	-.05

Table 14.10. (continued)

Predictor	Education Tier	REPETE Scale Scores Computer Courses	Level of Mastery	General Computer Skills	Basic Computer Certs.	Advanced Computer Certs.
RBI Scale Scores (cont.)						
Stress Tolerance	.07	**.11**	.04	.05	.01	.07
Hostility to Authority	.06	-.03	.05	.00	.08	.00
Self-Esteem	.07	-.01	.06	.07	.00	.05
Narcissism	-.01	**-.10**	.02	.03	-.03	-.08
Cultural Tolerance	**.10**	.01	.08	.07	-.05	-.04
Internal Locus of Control	.07	.05	.06	.04	-.05	-.02
Army Identification	**.18**	-.03	-.04	-.07	-.06	-.08
Respect for Authority	**-.11**	.02	.05	**.10**	-.05	-.02
Lie Scale	.04	-.08	-.07	**-.12**	-.04	-.01
WSI Scale Scores						
Achievement/Effort	-.04	.06	.08	.10	.02	.00
Adaptability/Flexibility	-.02	-.01	.02	.07	.00	.04
Attention to Detail	-.01	.02	.04	.02	.00	.02
Concern for Others	-.08	-.06	-.05	-.02	**-.11**	-.03
Cooperation	**-.08**	.08	-.06	-.01	.01	.04
Dependability	-.05	.00	.02	.05	-.05	-.01
Energy	.02	-.03	-.05	-.06	-.05	-.03
Independence	.01	-.04	-.05	-.05	.02	.06
Initiative	.07	-.05	.00	.01	.08	.03
Innovation	.04	.01	**.13**	**.10**	.04	.03
Leadership Orient	.07	-.01	.01	.01	-.07	**-.09**
Persistence	.03	-.08	.03	-.03	.01	.01
Self-Control	-.01	.03	-.03	-.06	.05	.02
Social Orientation	.03	.06	**-.11**	**-.09**	.03	-.03
Stress Tolerance	-.02	.02	.03	.02	.00	.02
Cultural Tolerance	.04	.00	.00	-.04	.03	-.06
WVI Composite Scores						
Growth	.00	.01	.04	.07	-.03	-.01
Status	.03	.00	.03	.06	.01	-.01
Stimulation	-.02	.02	.06	.05	-.01	-.04
Comfort	-.04	.02	-.01	-.02	.02	.01
Altruism	-.02	-.03	-.02	-.02	-.05	-.05
Self-Direction	.01	-.03	.04	.03	.00	.01
WPS Scale Scores						
Realistic	**.09**	-.07	**-.08**	**-.13**	-.04	-.06
Investigative	-.02	**.09**	**.21**	**.21**	-.02	.03
Artistic	.00	.03	**.15**	**.14**	**.11**	.06
Social	**-.08**	.01	.06	.05	-.05	-.07
Enterprising	-.05	.08	**.14**	**.13**	-.05	-.02
Conventional	-.05	.07	**.09**	**.11**	-.02	-.02

Table 14.10. (continued)

Predictor	Education Tier	REPETE Scale Scores Computer Courses	Level of Mastery	General Computer Skills	Basic Computer Certs.	Advanced Computer Certs.
IFQ Scale Scores						
Realistic	**.11**	-.02	-.06	-.05	-.02	.02
Investigative	.06	**.09**	**.19**	**.17**	.04	.06
Artistic	.01	-.01	**.11**	**.10**	.03	-.01
Social	.00	-.01	**.09**	.07	.00	-.03
Enterprising	-.02	.05	**.22**	**.19**	.03	-.03
Conventional	.04	.05	**.11**	**.13**	.04	.01

Note. n = 210 to 670. Certs. = Certifications. Bolded correlations are significant, $p < .05$ (two-tailed).

Temperament

Three Select21 predictors—the RBI, the WSI, and the PSJT (temperament scales)—were designed to measure personality-related variables. Correlations among scale scores from these instruments are shown in Tables 14.11 (RBI scales with WSI and PSJT scales) and 14.12 (WSI and PSJT scales). Correlations between RBI and WSI scores were rather modest. In fact, the highest correlation was .24 between RBI Peer Leadership and WSI Leadership Orientation. There was, however, some evidence for convergent validity. For example, in addition to the correlation between the two leadership scales, there were significant relations between RBI and WSI scale scores of similar constructs, including Fitness Motivation and Energy (r = .24), the two Cultural Tolerance scales (r = .23), Achievement and Achievement/Effort (r = .19), and Cognitive Flexibility and Innovation (r = .15).

Relations between RBI and PSJT temperament scale scores tended to be somewhat stronger and more consistent than those between RBI and WSI scores. However, many of the RBI-PSJT correlations varied by PSJT Form (A or B). Several logically related RBI and PSJT scale scores (Form A/Form B) were significantly correlated, including Achievement and Achievement Orientation (r = .21/.32), Achievement and Self-Efficacy (r = .25/.29), Hostility to Authority and Agreeableness (r = -.37/-.29), and Cultural Tolerance and Team Orientation (r = .23/.35). At the same time, however, there were correlations between RBI and PSJT scales that we would not necessarily expect to be related, such as Achievement and Agreeableness (r = .25/.24), Self-Efficacy and Sociability (r = .25/.16), and Locus of Control and Team Orientation (r = .23/.25).

As Table 14.12 shows, there were relatively few statistically significant correlations between WSI and PSJT temperament scale scores. Additionally, many of the correlations were significant for one PSJT form but not for the other and/or the WSI scales correlated with the two PSJT forms in different directions. Two factors that likely influence this difficult to interpret pattern of correlations are (a) the generally low reliability estimates (alpha) for the PSJT temperament scales (see Table 10.12), and (b) the ipsative format (and potential unreliability) of the WSI scale scores.

Table 14.11. Correlations among Temperament Predictor Measures (RBI Scales with WSI and PSJT Scales)

	Rational Biodata Inventory (RBI) Scale Scores													
Predictor	Peer Lead	Cog Flex	Achieve	Fitness Motive	Diplomacy	Stress Tolerance	Hostile to Authority	Self-Esteem	Narcissism	Cultural Tolerance	Internal LOC	Army ID	Respect Authority	Lie Scale
WSI Scale Scores														
Achievement/Effort	.04	**.09**	**.19**	**.12**	-.03	.01	-.02	**.14**	**.08**	.01	.06	**.10**	.03	.01
Adaptability/Flexibility	**-.09**	-.01	-.05	-.06	-.04	**.11**	-.01	-.05	**-.12**	-.01	-.01	-.03	-.01	-.04
Attention to Detail	.02	.01	**.14**	.08	-.07	-.01	-.05	.04	.03	.01	.00	.08	.03	.03
Concern for Others	-.05	-.01	-.05	**-.23**	-.02	-.08	**-.09**	**-.14**	-.01	.00	-.04	**-.18**	.07	.00
Cooperation	**-.13**	-.04	-.08	**-.10**	.03	-.02	.00	**-.14**	**-.09**	-.03	-.08	**-.12**	.02	.01
Dependability	.00	-.02	**.13**	.06	-.01	.02	-.06	-.01	.07	-.03	.08	**.13**	.02	.00
Energy	-.04	**-.12**	-.05	**.24**	-.05	-.08	.09	-.02	.01	-.07	-.04	.08	-.02	.04
Independence	-.04	-.05	**-.11**	**-.10**	**-.13**	-.08	-.02	-.01	-.01	**-.14**	-.01	**-.16**	-.06	.05
Initiative	.00	-.07	-.05	**-.12**	.01	.04	.01	.02	-.01	-.04	-.02	.08	-.04	.03
Innovation	**.12**	**.15**	**-.09**	-.05	.04	**-.12**	**.15**	-.01	.05	-.03	-.04	**-.11**	**-.11**	-.05
Leadership Orient	**.24**	.05	**.13**	.07	**.19**	.00	.07	**.13**	**.20**	.03	.06	**.13**	.01	-.01
Persistence	.00	**.13**	.08	.08	.02	.07	-.07	**.11**	-.03	.09	.02	.07	.03	**.09**
Self-Control	.01	.01	.00	.07	-.01	**.11**	-.03	.06	-.05	.04	.06	.08	-.06	-.02
Social Orientation	**-.09**	**-.15**	**-.09**	-.03	.07	.00	.06	-.08	-.04	-.03	-.04	-.04	.04	-.04
Stress Tolerance	.03	-.02	.02	**.14**	-.05	.07	-.03	.06	-.06	-.03	.05	**.11**	.01	-.03
Cultural Tolerance	-.04	.05	-.08	**-.14**	.07	.00	-.01	-.07	-.03	**.23**	-.03	**-.12**	.04	-.06
PSJT Scale Scores														
Achievement Orient	**.13/.25**	**.22/.28**	**.21/.32**	.07/.09	.08/**.26**	.01/.01	**-.23/-.29**	**.25/.29**	**.15/.21**	**.28/.21**	**.18/.28**	.13/**.15**	.05/.21	-.03/-.04
Self-Reliance	**.21**/.03	**.17**/.09	**.24**/.07	.09/.11	**.14**/-.04	-.03/-.02	**-.15**/-.03	**.24**/.12	**.15/.14**	**.28**/.03	**.24**/.03	.14/.04	-.05/-.03	.10/.01
Dependability	.06/**.17**	**.17/.28**	.11/**.32**	**.13**/.12	.09/**.22**	.03/.10	**-.21/-.47**	**.21/.33**	.00/**.19**	**.21/.31**	**.22/.36**	.13/**.24**	.08/.21	-.09/-.02
Sociability	**.18**/.10	**.23**/.09	**.19/.14**	-.01/.10	.06/.09	.04/-.06	**-.21**/.00	**.25/.16**	.10/.05	**.28**/.08	**.28**/.05	.15/.02	.10/.06	-.02/**-.20**
Agreeableness	.10/.09	**.16**/.12	**.25/.24**	.11/**.14**	**.19/.14**	.11/.05	**-.37/-.29**	**.24/.22**	**.15**/.10	**.32/.19**	**.19/.18**	.10/.09	.05/.15	.03/.04
Social Perceptiveness	**.20**/-.02	**.24**/.07	**.21/-.13**	.01/**-.16**	**.19**/-.01	.02/.03	**-.26**/-.02	**.25**/-.10	**.18**/-.06	**.31**/.05	**.25/.15**	.09/.04	.12/.05	.02/**-.19**
Team Orientation	.01/**.22**	.10/**.30**	**.18/.25**	.03/.11	.01/**.26**	.00/.07	**-.23/-.31**	.11/**.28**	.10/**.14**	**.24/.30**	**.23/.35**	.10/**.22**	.14/.19	-.03/-.05

Note. n = 236 to 565. Peer Lead = Peer Leadership. Cog Flex = Cognitive Flexibility. Achieve = Achievement. Fitness Motive = Fitness Motivation. Hostile to Authority = Hostility to Authority. Internal LOC = Internal Locus of Control. Army ID = Army Identification. Respect Authority = Respect for Authority. PSJT temperament scale scores are based only on items from the final version of the PSJT (12 items on Form A and 14 items on Form B). Correlations with PSJT Form A and Form B scale scores appear before and after the backslash, respectively. Bolded correlations are significant, $p < .05$ (two-tailed).

Table 14.12. Correlations among Temperament Predictor Measures (PSJT Temperament Scales with WSI Scales)

	Predictor Situational Judgment Test (PSJT) Temperament Scale Scores						
WSI Scale Scores	Achieve Orientation	Self-Reliance	Dependable	Sociability	Agreeable	Social Perception	Team Orientation
Achievement/Effort	.08/-.07	.04/-.03	.04/-.01	-.01/-.02	-.01/.00	.01/.00	.10/-.12
Adaptability/Flexibility	-.05/-.08	.00/**-.15**	.09/-.03	-.07/.00	-.10/-.06	-.06/.05	-.11/.01
Attention to Detail	.00/-.03	.04/.04	.02/.03	.07/-.07	**.14**/.04	.08/-.09	.06/.00
Concern for Others	-.07/.02	-.11/-.02	-.05/.03	-.10/.03	.04/.04	.01/.04	-.06/.07
Cooperation	-.10/.00	-.12/-.01	-.11/.02	-.06/-.08	**-.14**/.02	-.08/-.02	**-.14**/.05
Dependability	**.17**/-.05	.10/.01	.09/.00	**.18**/**-.17**	.08/-.06	.10/.09	**.22**/-.02
Energy	-.09/-.05	.03/.01	.05/**-.13**	.04/.05	-.01/-.06	-.06/-.03	**-.13**/-.04
Independence	.05/-.07	.04/.06	.01/.02	.01/.02	-.06/-.05	-.01/.11	.01/-.02
Initiative	.04/-.08	.11/-.08	.01/-.08	-.09/-.04	.03/-.01	.11/-.03	.09/-.06
Innovation	.11/-.01	.10/.01	-.02/-.04	.04/.03	-.03/-.06	.04/**.13**	.06/-.12
Leadership Orientation	-.02/.06	.00/.06	-.06/.00	-.09/.01	-.05/.02	.01/.06	-.04/.03
Persistence	.04/.06	.08/-.04	.03/.07	.04/-.06	-.05/-.07	.01/.07	.10/.05
Self-Control	.04/.11	.08/-.02	.02/.04	**.14**/.00	.09/.10	.01/-.03	-.03/.05
Social Orientation	-.05/.03	**-.25**/-.06	**-.15**/-.03	-.07/.09	.01/-.02	**-.13**/**-.17**	-.05/-.04
Stress Tolerance	-.07/.06	-.02/.05	.01/.06	.02/.12	.02/.11	-.05/-.10	-.05/**.13**
Cultural Tolerance	-.05/.10	-.08/**.15**	.04/.06	-.03/.05	.07/.07	.04/-.09	.00/.05

Note. $n = 241$. Achieve Orientation = Achievement Orientation. Dependable = Dependability. Agreeable = Agreeableness. Social Perception = Social Perceptiveness. PSJT temperament scale scores are based only on items from the final version of the PSJT (12 items on Form A and 14 items on Form B). Correlations with PSJT Form A and Form B scale scores appear before and after the backslash, respectively. Bolded correlations are significant, $p < .05$ (two-tailed).

Note, however, that like the PSJT Overall Judgment scores, the PSJT temperament scale scores used in these analyses are based on the final set of PSJT items (12 in Form A and 14 in Form B). For comparison, we also computed correlations between PSJT temperament scores based on the original PSJT instrument (32 items in each form) and scores from the other two temperament measures (see Table 14.13). In general, a consistent pattern of correlations (with the RBI and WSI) emerged between PSJT scores from the original and final forms. However, given the larger number of items on the original PSJT forms, it is not surprising that the magnitude of relations between the PSJT and RBI tended to be larger for temperament scores based on the original forms than for scores based on the shorter final forms (although original versus final PSJT form did not appear to influence PSJT-WSI correlations).

Correlations between the temperament measures and the other predictors are presented in Table 14.14. As discussed, several of the RBI scale scores were significantly related to the PSJT Overall Judgment scores, as well as to the REPETE Mean Level of Mastery and General Computer Skills scores. The RBI also demonstrated consistent, yet modest correlations with the WVI composite scores. Although not a temperament scale per se, RBI Army Affect had the strongest relationship with the WVI composites. These relations make sense, as recruits who hold values the work Army environment supports (e.g., Growth) are likely to have greater affect for the Army, whereas recruits who hold values the Army does not support (e.g., Comfort) are likely to have more negative feelings about the Army. Scores on the RBI were most related to vocational

interests scales of theoretically relevant constructs. The largest correlations were between RBI Cognitive Flexibility and the WPS and IFQ Investigative scale scores (r = .49 and .51). Cognitive Flexibility also correlated with the two Artistic scales (r = .36 and .33). These relationships make sense given that individuals high on Cognitive Flexibility like to try new approaches to solving problems (Investigative) and are willing to accept change and innovation (Artistic). Finally, it is sensible that RBI Peer Leadership and Achievement scale scores would be related to Enterprising interests (e.g., r = .28 and .34 with WPS Enterprising). Overall, these results provide some evidence of convergent validity for both the RBI and vocational interests measures.

Table 14.13. Correlations between PSJT Temperament Scores based on Original Instrument and Scores from the other Temperament Predictors

	Predictor Situational Judgment Test (PSJT) Temperament Scale Scores						
Predictor	Achieve Orientation	Self-Reliance	Dependable	Sociability	Agreeable	Social Perception	Team Orientation
RBI Scale Scores							
Peer Leadership	**.17/.27**	**.13/.17**	**.17/.24**	**.21/.20**	.09/.07	**.17/.23**	**.21/.26**
Cognitive Flexibility	**.21/.26**	.12/**.15**	**.24/.30**	**.28/.19**	**.22/.17**	**.25/.28**	**.22/.33**
Achievement Orientation	**.25/.35**	**.16/.17**	**.25/.32**	**.24/.30**	**.22/.17**	**.23/.27**	**.24/.33**
Fitness Motivation	.06/.12	.12/**.13**	.08/**.17**	.04/**.13**	-.01/.03	-.01/.02	.06/**.14**
Diplomacy	.09/**.26**	.02/.12	**.19/.23**	**.12/.24**	**.18**/.11	**.19/.22**	**.15/.28**
Stress Tolerance	-.05/.03	-.05/-.03	.04/.05	.01/.06	.15/.06	.12/-.05	-.03/.04
Hostility to Authority	**-.15/-.24**	.00/-.02	**-.31/-.38**	**-.14/-.17**	**-.41/-.39**	**-.42/-.20**	**-.22/-.28**
Self-Esteem	**.28/.31**	**.20/.22**	**.37/.35**	**.34/.28**	**.24/.18**	**.28/.27**	**.30/.33**
Narcissism	**.22/.24**	.10/**.16**	**.19/.18**	.11/.10	.10/-.04	.11/.22	**.20/.19**
Cultural Tolerance	**.26/.17**	**.14**/.09	**.35/.29**	**.29/.18**	**.37/.28**	**.41/.20**	**.31/.32**
Internal Locus of Control	**.14/.25**	.05/.01	**.25/.31**	**.20/.21**	**.16/.22**	**.30/.23**	**.27/.32**
Army Identification	**.15/.15**	**.15**/.00	**.22/.23**	**.18/.18**	.11/**.16**	.12/**.14**	**.26/.22**
Respect for Authority	.08/**.19**	-.01/-.03	.05/**.19**	.11/**.15**	.11/**.18**	.12/**.18**	.12/**.22**
Lie Scale	-.04/.04	.02/.06	.00/-.05	.04/-.09	.01/-.07	.06/.01	-.03/-.02
WSI Scale Scores							
Achievement/Effort	.10/-.02	.08/-.01	.06/-.01	.01/-.04	-.01/.07	-.01/.06	.03/-.02
Adaptability/Flexibility	-.01/-.10	.00/**-.13**	-.01/-.03	-.08/-.08	-.08/.04	-.08/-.02	-.06/-.05
Attention to Detail	-.01/-.02	.05/-.06	.07/.04	.04/-.06	.12/**.13**	.10/.01	.03/.02
Concern for Others	-.11/.04	**-.14**/-.07	-.03/.04	-.09/-.01	.06/.05	.05/.03	-.04/.01
Cooperation	-.12/-.01	-.05/-.05	-.09/.03	-.10/-.03	-.08/.05	-.10/.03	-.05/.01
Dependability	**.18**/-.06	.14/**-.14**	.10/-.02	.08/-.12	.02/.02	.06/-.03	**.18/-.06**
Energy	-.07/-.07	.10/.01	-.06/-.06	-.03/-.09	-.08/-.11	-.11/-.05	-.05/-.04
Independence	.07/-.07	.10/-.04	.03/.01	.00/.00	.03/-.06	.01/.05	.03/-.08
Initiative	.02/-.02	.02/-.01	.03/-.02	-.03/-.08	.07/-.07	.09/-.07	.04/-.07
Innovation	.09/-.01	.04/.04	.00/-.07	.05/-.01	-.01/-.06	.08/.08	.00/-.06
Leadership Orient	.04/.06	-.02/**.13**	.02/-.03	.04/.12	-.06/-.03	.01/.01	.07/.05
Persistence	.03/-.01	.12/.03	.01/.06	.11/-.01	-.04/-.01	-.04/.03	.03/.00
Self-Control	.02/.07	-.02/.02	.06/-.02	.11/.05	.06/.03	.06/-.03	-.01/.04
Social Orientation	-.06/.06	**-.15**/.07	-.12/.00	-.09/.09	-.05/-.04	-.12/-.07	.01/.01
Stress Tolerance	-.08/.06	-.12/**.13**	-.06/.08	-.02/**.14**	-.04/-.01	-.10/-.01	-.12/**.15**
Cultural Tolerance	-.07/.09	-.12/.08	.00/.01	.00/.10	.10/.01	.11/-.04	-.07/.08

Note. n = 241. Achieve Orientation = Achievement Orientation. Dependable = Dependability. Agreeable = Agreeableness. Social Perception = Social Perceptiveness. Correlations with PSJT Form A and Form B scale scores appear before and after the backslash. Bolded correlations are significant, $p < .05$ (two-tailed).

Table 14.14. Correlations between Temperament Predictors and Other Predictor Measures

	Rational Biodata Inventory (RBI) Scale Scores													
Predictor	Peer Lead	Cog Flex	Achieve	Fitness Motive	Diplomacy	Stress Tolerance	Hostile to Authority	Self-Esteem	Narcissism	Cultural Tolerance	Internal LOC	Army ID	Respect Authority	Lie Scale
ASVAB Composite Scores														
AFQT	**.16**	**.22**	**.09**	**.10**	.03	**.15**	**-.15**	.07	.00	**.10**	**.19**	.01	.07	**-.12**
Verbal	**.20**	**.25**	-.02	**.09**	.03	**.16**	**-.12**	.06	-.04	**.12**	**.17**	.08	.02	**-.13**
Quantitative	**.09**	**.15**	**.09**	.08	-.01	**.13**	**-.17**	.07	-.03	.06	**.14**	-.05	**.09**	-.07
Technical	**.08**	**.10**	-.08	**.13**	.01	**.19**	-.01	.08	**-.09**	.04	**.09**	**.11**	-.07	-.04
Spatial	.01	.06	-.01	.06	.00	.08	**-.12**	.02	-.01	**.12**	.07	.05	-.01	.01
Psychomotor Ability Test														
Time-to-Fire Scale	.04	.06	-.08	**.10**	.00	**.10**	-.01	-.01	-.03	.07	**.09**	.06	-.02	-.05
Precision Composite	.06	.08	-.03	**.10**	-.01	**.10**	-.02	.02	-.06	.02	.07	.07	-.02	-.02
PSJT Overall Judgment	**.16**	**.22**	**.24**	.06	**.21**	.06	**-.39**	**.24**	**.13**	**.34**	**.31**	**.19**	**.10**	**-.11**
Education Tier	.08	-.01	**-.08**	.01	**.10**	.07	.06	.07	-.01	**.10**	.07	**.18**	**-.11**	.04
REPETE Scale Scores														
Computer Courses Taken	.05	.07	.04	.08	-.01	**.11**	-.03	-.01	**-.10**	.01	.05	-.03	.02	-.08
Mean Level of Mastery	**.26**	**.25**	**.15**	**.15**	.05	.04	.05	.06	.02	.08	.06	-.04	.05	-.07
General Computer Skills	**.24**	**.22**	**.18**	**.13**	.07	.05	.00	.07	.03	.07	.04	-.07	**.10**	**-.12**
Basic Computer Certs.	.02	.00	.01	.01	-.01	.01	.08	.00	-.03	-.05	-.05	-.06	-.05	-.04
Advanced Computer Certs.	.02	.05	.01	.03	-.05	.07	.00	.05	-.08	-.04	-.02	-.08	-.02	-.01
WVI Composite Scores														
Growth	.04	**.13**	**.12**	.06	.06	**.13**	**-.16**	**.14**	.04	**.19**	**.14**	**.20**	**.15**	.06
Status	**.10**	.06	**.16**	.09	**.12**	-.01	-.06	.07	**.15**	**.12**	**.09**	**.16**	**.12**	-.02
Stimulation	.06	.04	.06	**.19**	.07	-.06	.02	.09	**.11**	**.12**	.01	**.19**	.07	.01
Comfort	-.03	-.07	**-.11**	-.05	.05	-.06	.05	-.09	.06	.00	-.02	**-.17**	-.04	-.07
Altruism	**.09**	**.14**	**.16**	**.10**	**.16**	.07	-.08	**.10**	.06	**.19**	**.12**	**.26**	**.19**	.01
Self-Direction	.09	**.12**	-.05	-.02	-.01	-.09	-.01	.05	**.10**	.06	.01	-.09	.03	-.01
WPS Scale Scores														
Realistic	-.06	**-.09**	-.04	**.24**	-.07	**-.11**	**.14**	.02	.03	-.06	-.07	**.20**	-.04	-.06
Investigative	**.26**	**.49**	**.33**	.07	**.12**	**.09**	**-.14**	**.23**	**.13**	**.20**	**.19**	**.10**	**.17**	.06
Artistic	**.18**	**.36**	.04	**-.12**	**.09**	**-.08**	.04	.03	**.08**	**.12**	-.04	**-.14**	.06	-.05
Social	**.20**	**.29**	**.29**	-.06	**.23**	-.05	**-.12**	**.14**	**.13**	**.21**	**.12**	.04	**.20**	.03
Enterprising	**.28**	**.30**	**.34**	.04	**.21**	-.03	-.06	**.18**	**.26**	**.14**	**.12**	**.12**	**.15**	.05
Conventional	.02	**.19**	**.28**	-.04	.02	.00	**-.11**	**.08**	**.08**	.07	.06	.01	**.17**	.06
IFQ Scale Scores														
Realistic	-.04	-.03	-.06	**.13**	**-.09**	-.08	**.12**	.02	-.03	-.05	-.06	.06	-.02	-.06
Investigative	**.21**	**.51**	**.19**	**.11**	.07	**.13**	**-.12**	**.21**	.08	**.24**	**.15**	**.13**	**.14**	.02
Artistic	**.20**	**.33**	.06	-.07	**.14**	**-.11**	.01	**.10**	**.12**	**.17**	.04	.00	**.12**	**-.08**
Social	**.23**	**.27**	**.25**	-.06	**.20**	-.05	-.08	**.15**	**.14**	**.18**	**.13**	.03	**.16**	-.04
Enterprising	**.34**	**.31**	**.25**	**.11**	**.27**	.00	**.13**	**.22**	**.23**	**.17**	.07	**.08**	**.10**	-.05
Conventional	-.04	**.10**	**.14**	**-.12**	.03	-.01	**-.11**	.02	.03	.05	.08	-.08	**.12**	.06

Table 14.14. (continued)

	Work Suitability Inventory (WSI) Scale Scores															
Predictor	Achieve/ Effort	Adapt/ Flexible	Attend Details	Concern Others	Cooperate	Depend	Energy	Indepndt	Initiative	Innovate	Leader Orient	Persist	Self-Control	Social Orient	Stress Toleranc	Cultural Tolerance
ASVAB Composite Scores																
AFQT	.00	-.03	.02	**-.10**	**-.23**	.08	.04	.08	.00	.05	.05	.06	**.12**	**-.12**	.08	-.09
Verbal	-.08	-.03	.00	**-.10**	**-.20**	.01	-.01	**.10**	.02	**.09**	.04	.04	**.14**	-.06	**.11**	-.06
Quantitative	.03	**-.08**	.01	**-.09**	**-.16**	.07	.00	.05	.00	.04	.07	**.09**	**.08**	**-.10**	.06	-.05
Technical	**-.08**	-.08	.05	**-.23**	**-.23**	-.01	.05	**.14**	.04	**.10**	.05	**.17**	**.16**	**-.08**	**.10**	**-.12**
Spatial	-.05	.00	-.02	**-.17**	**-.12**	.11	.01	.05	-.03	.05	**.10**	**.10**	.06	-.05	.01	-.02
Psychomotor Ability Test																
Time-to-Fire Scale	**-.08**	.01	-.03	-.05	**-.10**	-.07	.00	.03	-.08	.02	.05	.07	.05	.04	.06	.05
Precision Composite	**-.09**	-.04	-.02	**-.11**	**-.10**	-.01	.01	.01	-.04	-.02	.04	.07	**.11**	.07	**.12**	-.01
PSJT Overall Judgment	-.02	.00	.02	-.01	**-.11**	**.10**	-.04	.02	-.02	.00	.00	.02	**.10**	-.07	.01	.02
Education Tier	-.04	-.02	-.01	-.08	**-.08**	-.05	.02	.01	.07	.04	.07	.03	-.01	.03	-.02	.04
REPETE Scale Scores																
Computer Courses Taken	.06	-.01	.02	-.06	.08	.00	-.03	-.04	-.05	.01	-.01	-.08	.03	.06	.02	.00
Mean Level of Mastery	.08	.02	.04	-.05	-.06	.02	-.05	-.05	.00	**.13**	.01	.03	-.03	**-.11**	.03	.00
General Computer Skills	**.10**	.07	.02	-.02	-.01	.05	-.06	-.05	.01	**.10**	.01	-.03	-.06	**-.09**	.02	-.04
Basic Computer Certs.	.02	.00	.00	**-.11**	.01	-.05	-.05	.02	.08	.04	-.07	.01	.05	.03	.00	.03
Advanced Comp Certs.	.00	.04	.02	-.03	.04	-.01	-.03	.06	.03	.03	**-.09**	.01	.02	-.03	.02	-.06
WVI Composite Scores																
Growth	.08	-.09	.02	**-.09**	-.06	**.12**	**.10**	**-.18**	.08	.00	.04	.07	.01	-.05	.06	-.07
Status	.02	**-.12**	.02	-.01	-.04	.08	.01	**-.22**	.00	-.01	**.09**	.04	.09	.04	.04	.00
Stimulation	-.05	-.01	-.03	**-.10**	-.01	-.05	**.23**	**-.15**	.00	-.01	-.03	-.04	.04	.06	**.14**	.02
Comfort	**-.25**	.00	**-.17**	**.22**	**.11**	-.08	-.03	-.03	**-.11**	**.11**	.00	-.10	.03	**.14**	**-.12**	**.19**
Altruism	.07	-.08	.00	.08	.05	.02	.03	**-.35**	.00	-.06	**.11**	-.01	-.02	**.14**	.00	.06
Self-Direction	**-.10**	-.01	-.08	.02	-.05	-.04	-.04	**.22**	-.04	**.16**	.00	.02	.00	**-.14**	-.03	.05
WPS Scale Scores																
Realistic	.07	-.07	.06	**-.26**	**-.11**	**.08**	**.24**	-.04	.08	-.05	-.02	**.09**	.03	-.02	**.13**	**-.14**
Investigative	**.19**	-.01	**.11**	-.04	**-.10**	-.02	**-.13**	**-.14**	.07	.02	.04	**.13**	.03	**-.12**	.00	.01
Artistic	-.08	.02	**-.12**	**.14**	**.11**	**-.12**	**-.16**	-.01	-.03	**.38**	-.06	-.01	-.02	-.08	**-.16**	**.16**
Social	.07	.00	-.01	**.26**	**.12**	-.07	**-.11**	**-.26**	.05	-.07	.04	-.05	-.06	.08	-.08	**.11**
Enterprising	**.14**	-.07	.05	.00	-.03	-.03	**-.09**	**-.14**	**.10**	-.08	**.24**	.03	-.03	-.06	.02	-.02
Conventional	**.18**	.01	**.17**	.05	.05	.07	**-.13**	**-.14**	**.10**	**-.15**	-.06	.01	-.07	-.07	-.04	.03
IFQ Scale Scores																
Realistic	.02	**-.10**	.02	**-.17**	-.05	**.08**	**.12**	.04	.02	.03	-.03	.06	.04	-.05	.06	-.07
Investigative	.03	-.02	.04	-.07	**-.14**	-.04	-.07	-.03	-.04	**.12**	.06	**.12**	.07	**-.09**	.00	.07
Artistic	-.07	-.02	**-.12**	**.10**	.03	-.06	-.06	**-.09**	-.02	**.26**	.01	-.07	.00	.00	-.06	**.14**
Social	.01	.02	-.03	**.19**	.07	-.08	**-.08**	**-.12**	.01	.01	.05	**-.10**	-.03	.01	-.08	**.11**
Enterprising	.03	-.06	-.06	-.03	-.04	-.05	**-.08**	**-.10**	.03	**.09**	**.20**	.00	.04	-.05	.02	.08
Conventional	.06	-.01	.04	**.12**	**.11**	**.08**	**-.19**	-.03	**.08**	-.06	-.06	-.04	-.08	-.05	-.06	**.09**

Table 14.14. (continued)

Predictor	Predictor Situational Judgment Test (PSJT) Temperament Scale Scores						
	Achieve Orientation	Self-Reliance	Dependable	Sociability	Agreeable	Social Perception	Team Orientation
ASVAB Composite Scores							
AFQT	-.02/**.19**	.05/-.01	.09/**.15**	.06/.04	.05/-.03	.01/**.27**	-.05/**.18**
Verbal	.02/**.21**	.08/.06	.09/.08	**.11**/.10	.01/-.03	.01/**.26**	-.01/**.18**
Quantitative	-.06/**.14**	.02/-.01	.05/**.16**	-.01/-.04	.05/-.02	-.01/**.21**	-.07/**.13**
Technical	-.04/**.13**	.09/**.13**	.16/.02	.03/**.13**	.01/.03	-.02/.08	.01/.01
Spatial	-.01/.02	.09/-.02	.12/.08	.10/-.12	.05/-.08	-.03/**.16**	.05/.09
Psychomotor Ability Test							
Time-to-Fire Scale	-.09/.08	.04/.12	.04/.02	.08/-.04	.00/-.10	-.10/.02	-.08/-.03
Precision Composite	.00/.00	.06/.02	-.01/-.05	.08/-.07	.08/**-.18**	.01/.02	.01/-.01
PSJT Overall Judgment	**.68/.65**	**.57**/.03	**.51/.75**	**.52/.18**	**.60/.42**	**.66/.25**	**.63/.74**
Education Tier	-.02/.05	.06/.02	-.01/.10	.00/.01	.08/.08	.02/-.01	.05/.00
REPETE Scale Scores							
Computer Courses Taken	-.06/-.06	.05/.07	-.10/.01	-.01/-.09	-.09/-.01	**-.15**/-.08	-.08/-.08
Mean Level of Mastery	-.05/-.02	.01/.10	-.12/.10	.02/.03	-.08/.02	-.13/.02	-.12/.03
General Computer Skills	-.04/-.02	.02/.00	-.09/.07	.02/.01	-.04/.05	-.06/.00	-.08/.01
Basic Computer Certs.	-.05/-.13	**-.14**/-.05	-.11/-.12	-.01/-.12	.07/-.13	-.03/.02	-.09/**-.21**
Advanced Computer Certs.	-.06/-.03	**-.14**/.01	-.09/-.02	.02/-.07	.04/-.05	.00/-.01	-.11/-.07
WVI Composite Scores							
Growth	.03/.05	.03/.01	**.14**/.09	.11/-.02	.09/.00	.08/.12	.05/.05
Status	**.09**/.10	.05/.04	.10/.07	.04/-.05	.06/.02	.00/.09	.07/.06
Stimulation	.04/.01	-.02/.04	.12/-.02	.11/-.04	.01/-.07	.00/.05	-.11/.05
Comfort	-.02/.03	-.08/.07	.04/-.07	.05/.06	.01/-.03	-.04/.13	-.06/-.06
Altruism	.10/.04	.07/-.04	**.16**/.05	.11/.00	**.16**/.05	**.14/.06**	.00/.07
Self-Direction	-.01/.02	-.07/.11	.03/.02	.08/-.05	-.02/-.04	.03/.09	-.06/-.01
WPS Scale Scores							
Realistic	.09/-.04	**.16**/.07	**.18**/-.05	.02/.06	-.01/.09	.04/-.10	.06/-.08
Investigative	**.30/.23**	**.21**/.11	**.24/.17**	**.22**/.10	**.15/.18**	**.26**/-.07	**.17/.14**
Artistic	**.14/.20**	.12/.11	.10/.06	**.13/.14**	.05/.09	.05/.08	.05/.09
Social	**.28/.22**	**.24**/.03	**.13/.17**	**.17/.14**	.11/**.21**	**.29**/-.05	**.19/.17**
Enterprising	**.31/.20**	**.28**/.10	**.17/.13**	**.16**/.09	.06/**.19**	**.27**/-.11	**.17**/.11
Conventional	**.27**/.06	**.19**/-.02	.12/**.15**	.07/.07	**.18/.23**	**.20/-.13**	**.19**/.07
IFQ Scale Scores							
Realistic	.04/-.06	.10/.03	.09/-.09	-.02/.07	.06/.08	.03/.03	.02/-.05
Investigative	**.25/.17**	.11/.03	**.17/.13**	**.18**/.07	**.19**/.05	**.17**/.07	**.14/.20**
Artistic	**.15**/.12	.11/.05	.04/.08	**.14**/.12	.08/.05	.06/.08	.06/**.19**
Social	**.33/.17**	**.19**/.03	**.14/.17**	**.16**/.09	**.21/.19**	**.28**/.03	.08/**.24**
Enterprising	**.19**/.11	.09/.10	.05/.10	.10/.10	-.01/.12	.10/-.05	-.06/.10
Conventional	**.18**/-.02	.08/-.04	.05/**.18**	.03/-.02	**.18/.16**	**.19**/-.05	.10/.12

Note. n = 196 to 623. Peer Lead = Peer Leadership. Cog Flex = Cognitive Flexibility. Achieve = Achievement. Fitness Motive = Fitness Motivation. Hostile to Authority = Hostility to Authority. Internal LOC = Internal Locus of Control. Respect Authority = Respect for Authority. Achieve/Effort = Achievement/Effort. Adapt/Flexible = Adaptability/Flexibility. Attend Details = Attend to Details. Concern Others = Concern for Others. Cooperate = Cooperation. Depend = Dependability. Indepndt = Independence. Innovate = Innovation. Leader Orient = Leadership Orientation. Persist = Persistence. Social Orient = Social Orientation. Achieve Orientation = Achievement Orientation. Dependable = Dependability. Agreeable = Agreeableness. Social Perception = Social Perceptiveness. PSJT temperament scale scores are based only on items from the final version of the PSJT (12 items on Form A and 14 items on Form B). Correlations with PSJT Form A and Form B scale scores appear before and after the backslash, respectively. Bolded correlations are significant, $p < .05$ (two-tailed).

Correlations between the WSI scale scores and the other predictors were much smaller than the RBI correlations. Interestingly, WSI Concern for Others and Cooperation scores were negatively related to cognitive ability (e.g., $rs = -.23$ with ASVAB Technical composite scores). WSI scores were largely unrelated to psychomotor ability, judgment, and education, training, and experience, but did have some relationship to values and interests. For instance, WSI Independence was, in general, negatively correlated with the WVI composite scores. The single largest correlation between the WSI and the other predictors was .38 between Innovation and WPS Artistic. Innovation was also related to IFQ Artistic scores ($r = .26$). Other logical relationships included Concern for Others and WPS Social ($r = .26$) and Independence and WPS Social ($r = -.26$). There were additional, theoretically meaningful relations (e.g., between Cultural Tolerance and Artistic), but the magnitude of these correlations was notably smaller.

Finally, relations between the PSJT temperament scale scores (based on the final test forms) and the remaining predictor measures were also modest, and the nature of many of the relationships varied by PSJT form. PSJT scores correlated most consistently with the vocational interest scale scores, however; there does not appear to be a theoretical reason for many of these relationships. For instance, the single largest correlation was .33 between PSJT Achievement Orientation (Form A) and IFQ Social. There was, however, some evidence of convergent validity. As an example, PSJT Agreeableness and Team Orientation scores were positively correlated with WPS and IFQ Social scores. Nonetheless, there appears to be a lack of discriminant validity evidence given that these PSJT scales were similarly correlated with Conventional scale scores, which we would not expect to be related to Agreeableness and Team Orientation.

P-E Fit

The P-E fit predictors were designed to measure "fit" between recruits' work values and interests and the values/interests supported by the Army work environment. Table 14.15 presents correlations between the WVI scale and composite scores and the WPS and IFQ scale scores. Several expected relationships emerged from this analysis. For instance, WVI Societal Contribution and Social Service scores correlated with WPS and IFQ Social scores (e.g., $rs = .30$ with the IFQ). In addition, WVI Leadership Opportunities scores were related to WPS and IFQ Social and Enterprising scores (e.g., $r = .23$ and .33 with WPS scores), whereas WVI Physical Development scores correlated with WPS and IFQ Realistic scores ($r = .34$ and .14). Lastly, WVI Creativity scores correlated .21 and .16 with WPS and IFQ Artistic scores.

As for relations between the WVI composites and the WPS and IFQ scales, Altruism correlated the highest with the vocational interests scales. It is perhaps not surprising that the strongest relationship was between Altruism and WPS and IFQ Social scores ($r = .32$ and .29), as individuals with Social interests tend to like human services vocations and activities. Taken together, although the magnitude of relations between recruits' values and interests was generally quite modest (the largest correlation was .36 between WVI Social Service and WPS Social), the overall pattern of relationships provides some evidence for the construct-related validity of these measures.

Table 14.15. Correlations among P-E Fit Needs Predictor Measures

	WPS Scale Scores						IFQ Scale Scores					
Predictor	R	I	A	S	E	C	R	I	A	S	E	C
WVI Scale Scores												
Social Status	.08	.07	-.05	.06	**.20**	.06	-.02	.03	-.01	.04	.09	-.02
Advancement	.04	-.03	-.03	-.02	**.10**	.00	.05	.01	-.01	.06	**.19**	.07
Autonomy	-.03	-.05	.01	-.02	.00	-.04	-.03	.00	.04	.03	.07	-.02
Supportive Supervision	.08	.01	.01	**.15**	.02	**.17**	**.10**	.06	.03	**.14**	-.01	**.14**
Leisure Time	-.07	**-.15**	.03	**-.13**	**-.10**	**-.22**	-.01	-.06	.00	**-.10**	-.01	**-.17**
Comfort	-.08	**-.14**	**.09**	.02	-.08	.03	-.03	-.04	**.09**	.05	.03	.08
Achievement	-.05	**.13**	.08	**.10**	**.15**	.08	-.07	.03	.06	.08	.08	.02
Societal Contribution	.05	**.21**	.03	**.23**	**.18**	**.16**	.04	**.16**	**.12**	**.30**	**.14**	**.10**
Independence	-.03	-.01	.02	**-.11**	-.03	.06	-.01	.01	.03	.02	-.03	.08
Social Service	.02	**.15**	.05	**.36**	**.17**	**.14**	.00	.06	**.10**	**.30**	**.10**	**.09**
Fixed Role	.04	.04	-.04	**.09**	.04	**.20**	**.11**	.01	.02	.06	-.01	**.14**
Variety	**.09**	-.03	.01	.03	.01	-.02	.00	-.01	.04	.04	-.02	-.01
Leadership Opportunities	**.10**	**.17**	.03	**.23**	**.33**	**.11**	.04	.08	.07	**.17**	**.22**	-.02
Feedback	.07	**.09**	.01	.07	.09	**.09**	.04	.03	-.04	.06	.02	.07
Travel	.07	**.10**	**.10**	.06	**.11**	-.01	-.06	.06	.06	.06	.06	-.01
Physical Development	**.34**	.01	-.03	.01	.07	-.06	**.14**	.02	-.03	-.03	.02	**-.13**
Ability Utilization	.04	.06	.04	-.03	**-.09**	-.01	.00	.07	.03	-.04	-.02	-.04
Creativity	-.03	.04	**.21**	.00	-.03	-.06	-.03	.08	**.16**	-.02	.02	**-.09**
Recognition	-.01	-.02	.04	-.03	**.12**	.00	.00	.00	-.02	-.04	.08	.01
Co-Workers	.02	-.04	.00	**.15**	.01	-.03	.05	.01	.05	**.09**	.03	.02
Activity	**.09**	.00	-.07	.02	-.06	**.16**	.05	-.03	-.05	.03	-.07	**.14**
Flexible Schedule	**-.09**	**-.12**	**.11**	-.08	**-.12**	**-.13**	.00	-.04	.06	.02	.03	.00
Skill Development	**.10**	**.11**	.02	.05	.01	.08	.04	**.17**	-.02	.04	.03	.03
Home	-.04	-.08	.03	.01	-.05	-.01	.04	-.04	.01	.06	.03	.01
Esteem	**-.09**	**.10**	-.04	.02	**.10**	.07	**-.09**	.06	-.02	.00	.06	.03
Emotional Development	**.10**	.03	**-.10**	.01	.06	.02	.01	.00	-.07	-.02	-.06	-.03
Influence	-.03	**.14**	-.03	**.15**	**.20**	**.14**	-.02	.02	.03	**.09**	**.10**	.08
Team Orientation	-.03	.01	-.04	**.18**	.00	.05	-.02	-.01	.01	**.11**	.04	.06
WVI Composite Scores												
Growth	**.09**	**.10**	-.05	.06	.03	**.11**	.03	.07	-.03	.03	-.01	.05
Status	.04	.07	.00	.07	**.17**	**.10**	.03	.02	.00	.06	**.11**	.07
Stimulation	**.22**	.04	.04	.04	**.09**	-.04	.03	.04	.03	.03	.03	-.06
Comfort	-.08	**-.15**	.08	-.01	**-.10**	**-.11**	.02	-.05	.06	.03	.03	-.02
Altruism	.06	**.16**	.02	**.32**	**.20**	**.18**	.05	.10	**.09**	**.29**	**.13**	**.10**
Self-Direction	-.04	-.01	**.10**	-.06	-.03	-.01	-.03	.04	**.10**	.01	.02	-.01

Note. n = 487 to 523. R = Realistic. I = Investigative. A = Artistic. S = Social. E = Enterprising. C = Conventional. Bolded correlations are significant, $p < .05$ (two-tailed).

Correlations between the P-E fit needs measures and the remaining predictors are presented in Table 14.16.[91] In general, correlations between the WVI composites and scores from the other predictors were rather modest. In fact, the WVI was virtually unrelated to measures of ability, judgment, and education, training, and experience.[92] There were, however, some statistically significant correlations between WVI scores and the temperament measures. The largest correlation was .26 between the Altruism composite and RBI Army Affect. The WVI also demonstrated some relationship to WSI scale scores, such as correlations between Comfort composite scores and WSI Achievement/Effort ($r = -.25$) and Concern for Others ($r = .22$) scale scores.

Numerous statistically significant correlations were found between interests scale scores and the other predictor scores. Interestingly, WPS and IFQ Realistic scores were positively related to ASVAB Technical composite scores ($r = .27$ and .30). Further analysis revealed that these relations were due primarily to the correspondence between Realistic interests and the Auto and Shop Information test scores that contribute to the Technical composite. In addition, Investigative scores tended to correlate positively with cognitive ability scores (which makes sense given the nature of Investigative interests), whereas the remaining interests scores (particularly Social and Conventional scores) were negatively related to cognitive ability. Also, as discussed earlier, there were several significant correlations between interests and REPETE scores, with most scales (and Investigative most notably) relating positively to the REPETE. In contrast, Realistic scores tended to relate negatively to the REPETE scale scores, which is sensible given that Realistic individuals tend to not be interested in educational activities. Finally, with the exception of Realistic, interests scale scores correlated positively with PSJT Overall Judgment scores.

Expectations about the Army

The Army Beliefs Survey (ABS), Pre-Service Expectations Survey (PSES), and Army Work Knowledge Survey (AWKS) assess the extent to which recruits expect the Army to support work values, vocational interests, and temperament-related work activities, respectively. Correlations among scores from these measures are displayed in Table 14.17. ABS composite scores were moderately related to several PSES and AWKS scale scores. It is not surprising, for example, that Supported Reinforcers and Unsupported Reinforcers composites tended to correlate significantly with scales measuring interests and personality variables the Army work environment supports and does not support, respectively. For instance, the Supported composite correlated positively with PSES Realistic, Enterprising, and Conventional scores, and with AWKS Adaptability/Flexibility, Attention to Detail, and Dependability scores. Conversely, the Unsupported composite covaried with PSES Artistic and Investigative scores, and with AWKS Innovation and Concern for Others scores. Interestingly, the ABS Expects Recognition and Achievement composite correlated significantly with all PSES and AWKS scales except Independence. The correlation of .50 between ABS Supported Reinforcers and AWKS Energy was the largest among the expectations measures.

[91] Although this table contains correlations between P-E fit and temperament measures, relations between scores on these measures were discussed in the temperament section (see page 14-11) and thus are not reiterated here.
[92] One likely contributing factor to the modest relations between the WVI and the remaining predictors is the relatively low reliability estimates for some of the WVI composites (see Table 13.24).

Table 14.16. Correlations between P-E Fit Needs Predictors and the Other Predictor Measures

	WVI Composite Scores						WPS Scale Scores						IFQ Scale Scores					
Predictor	Growth	Status	Stimulate	Comfort	Altruism	Self-Direct	R	I	A	S	E	C	R	I	A	S	E	C
ASVAB Composite Scores																		
AFQT	.03	-.03	-.02	.02	**-.10**	.03	**-.10**	**.11**	-.05	**-.11**	-.01	**-.12**	-.02	**.16**	-.02	-.06	.06	**-.11**
Verbal	-.01	-.05	.00	.00	**-.10**	.05	-.03	**.09**	-.01	**-.12**	-.05	**-.27**	-.01	**.24**	.04	**-.08**	-.02	**-.28**
Quantitative	.03	-.07	-.05	-.01	**-.11**	.00	**-.14**	.08	-.06	**-.10**	-.03	-.02	-.04	**.11**	-.05	**-.09**	.03	.00
Technical	-.03	-.03	.00	.00	**-.12**	.03	**.27**	.03	-.03	**-.25**	**-.14**	**-.26**	**.30**	**.17**	**-.10**	**-.26**	**-.10**	**-.29**
Spatial	.06	.01	.04	-.05	-.07	-.03	.06	.00	.01	**-.11**	-.08	-.05	.07	.05	-.03	**-.15**	**-.09**	-.03
Psychomotor Ability Test																		
Time-to-Fire Scale	-.03	-.02	.00	.07	-.07	.00	.07	.04	-.04	**-.13**	-.05	**-.17**	-.01	**.09**	**-.08**	**-.16**	-.07	**-.17**
Precision Composite	.01	.00	.05	-.03	-.03	.00	**.11**	.07	-.03	**-.13**	-.04	**-.14**	.09	**.17**	-.04	**-.17**	-.01	**-.18**
PSJT Overall Judgment	**.14**	.09	.01	.01	**.11**	.01	-.03	**.20**	.05	**.19**	**.16**	**.13**	-.01	**.20**	**.13**	**.23**	.05	**.10**
Education Tier	.00	.03	-.02	-.04	-.02	.01	**.09**	-.02	.00	**-.08**	-.05	-.05	**.11**	.06	.01	.00	-.02	.04
REPETE Scale Scores																		
Computer Courses Taken	.01	.00	.02	.02	-.03	-.03	-.07	**.09**	.03	.01	.08	.07	-.02	**.09**	-.01	-.01	.05	.05
Mean Level of Mastery	.04	.03	.06	-.01	-.02	.04	**-.08**	**.21**	**.15**	.06	**.14**	**.09**	-.06	**.19**	**.11**	**.09**	**.22**	**.11**
General Computer Skills	.07	.06	.05	-.02	-.02	.03	**-.13**	**.21**	**.14**	.05	**.13**	**.11**	-.05	**.17**	**.10**	.07	**.19**	**.13**
Basic Computer Certs.	-.03	.01	-.01	.02	-.05	.00	-.04	-.02	**.11**	-.05	-.05	-.02	-.02	.04	.03	.00	.03	.04
Advanced Computer Certs.	-.01	-.01	-.04	.01	-.05	.01	-.06	.03	.06	-.07	-.02	-.02	.02	.06	-.01	-.03	-.03	.01
RBI Scale Scores																		
Peer Leadership	.04	**.10**	.06	-.03	**.09**	.09	-.06	**.26**	**.18**	**.20**	**.28**	.02	-.04	**.21**	**.20**	**.23**	**.34**	-.04
Cognitive Flexibility	**.13**	.06	.04	-.07	**.14**	**.12**	**-.09**	**.49**	**.36**	**.29**	**.30**	**.19**	-.03	**.51**	**.33**	**.27**	**.31**	.10
Achievement Orientation	**.12**	**.16**	.06	**-.11**	**.16**	-.05	-.04	**.33**	.04	**.29**	**.34**	**.28**	-.06	**.19**	.06	**.25**	**.25**	**.14**
Fitness Motivation	.06	.09	**.19**	-.05	**.10**	-.02	**.24**	.07	**-.12**	-.06	.04	-.04	**.13**	**.11**	-.07	-.06	**.11**	**-.12**
Diplomacy	.06	**.12**	.07	.05	**.16**	-.01	-.07	**.12**	**.09**	**.23**	**.21**	.02	**-.09**	.07	**.14**	**.20**	**.27**	.03
Stress Tolerance	**.13**	-.01	-.06	-.06	.07	-.09	**-.11**	**.09**	**-.08**	-.05	-.03	.00	-.08	**.13**	**-.11**	-.05	.00	-.01
Hostility to Authority	**-.16**	-.06	.02	.05	-.08	-.01	**.14**	**-.14**	.04	**-.12**	-.06	**-.11**	**.12**	**-.12**	.01	-.08	**.13**	**-.11**
Self-Esteem	**.14**	.07	.09	-.09	**.10**	.05	.02	**.23**	.03	**.14**	**.18**	**.08**	.02	**.21**	**.10**	**.15**	**.22**	.02
Narcissism	.04	**.15**	**.11**	.06	**.06**	**.10**	.03	**.13**	**.08**	**.13**	**.26**	**.08**	-.03	.08	**.12**	**.14**	**.23**	.03
Cultural Tolerance	**.19**	**.12**	**.12**	.00	**.19**	.06	-.06	**.20**	**.12**	**.21**	**.14**	.07	-.05	**.24**	**.17**	**.18**	**.17**	.05
Internal Locus of Control	**.14**	**.09**	.01	-.02	**.12**	.01	-.07	**.19**	-.04	**.12**	**.12**	.06	-.06	**.15**	.04	**.13**	.07	.08
Army Identification	**.20**	**.16**	**.19**	**-.17**	**.26**	-.09	**.20**	**.10**	**-.14**	.04	**.12**	.01	.06	**.13**	.00	.03	**.08**	-.08
Respect for Authority	**.15**	**.12**	.07	-.04	**.19**	.03	-.04	**.17**	.06	**.20**	**.15**	**.17**	-.02	**.14**	**.12**	**.16**	**.10**	**.12**
Lie Scale	.06	-.02	.01	-.07	.01	-.01	-.06	.06	-.05	.03	.05	.06	-.06	.02	**-.08**	-.04	-.05	.06

Table 14.16. (continued)

	WVI Composite Scores						WPS Scale Scores						IFQ Scale Scores					
Predictor	Growth	Status	Stimulate	Comfort	Altruism	Self-Direct	R	I	A	S	E	C	R	I	A	S	E	C
WSI Scale Scores																		
Achievement/Effort	.08	.02	-.05	**-.25**	.07	**-.10**	.07	**.19**	-.08	.07	**.14**	**.18**	.02	.03	-.07	.01	.03	.06
Adaptability/Flexibility	-.09	**-.12**	-.01	.00	-.08	-.01	-.07	-.01	.02	.00	-.07	.01	**-.10**	-.02	-.02	.02	-.06	-.01
Attention to Detail	.02	.02	-.03	**-.17**	.00	-.08	.06	**.11**	**-.12**	-.01	.05	**.17**	.02	.04	**-.12**	-.03	-.06	.04
Concern for Others	**-.09**	-.01	**-.10**	**.22**	.08	.02	**-.26**	-.04	**.14**	**.26**	.00	.05	**-.17**	-.07	**.10**	**.19**	-.03	**.12**
Cooperation	-.06	-.04	-.01	**.11**	.05	-.05	**-.11**	**-.10**	**.11**	**.12**	-.03	.05	-.05	**-.14**	.03	.07	-.04	**.11**
Dependability	**.12**	.08	-.05	-.08	.02	-.04	**.08**	-.02	**-.12**	-.07	-.03	.07	**.08**	-.04	-.06	-.08	-.05	.08
Energy	**.10**	.01	**.23**	-.03	.03	-.04	**.24**	**-.13**	**-.16**	**-.11**	**-.09**	**-.13**	**.12**	-.07	-.06	**-.08**	**-.08**	**-.19**
Independence	**-.18**	**-.22**	**-.15**	-.03	**-.35**	**.22**	-.04	**-.14**	-.01	**-.26**	**-.14**	**-.14**	.04	-.03	**-.09**	**-.12**	**-.10**	-.03
Initiative	.08	.00	.00	**-.11**	.00	-.04	.08	.07	-.03	.05	**.10**	**.10**	.02	-.04	-.02	.01	.03	**.08**
Innovation	.00	-.01	-.01	**.11**	-.06	**.16**	-.05	.02	**.38**	-.07	-.08	**-.15**	.03	**.12**	**.26**	.01	**.09**	-.06
Leadership Orient	.04	**.09**	-.03	.00	**.11**	.00	-.02	.04	-.06	.04	**.24**	-.06	-.03	.06	.01	.05	**.20**	-.06
Persistence	.07	.04	-.04	**-.10**	-.01	.02	**.09**	**.13**	-.01	-.05	.03	.01	.06	**.12**	-.07	**-.10**	.00	-.04
Self-Control	.01	**.09**	.04	.03	-.02	.00	.03	.03	-.02	-.06	-.03	-.07	.04	.07	.00	-.03	.04	-.08
Social Orientation	-.05	.04	.06	**.14**	**.14**	**-.14**	-.02	**-.12**	-.08	.08	-.06	-.07	-.05	**-.09**	.00	.01	-.05	-.05
Stress Tolerance	.06	.04	**.14**	**-.12**	.00	-.03	**.13**	.00	**-.16**	-.08	.02	-.04	.06	.00	-.06	-.08	.02	-.06
Cultural Tolerance	-.07	.00	.02	**.19**	.06	.05	**-.14**	.01	**.16**	**.11**	-.02	.03	-.07	.07	**.14**	**.11**	.08	**.09**

Note. n = 402 to 642. Stimulate = Stimulation. Self-Direct = Self-Direction. R = Realistic. I = Investigative. A = Artistic. S = Social. E = Enterprising. C = Conventional. Bolded correlations are significant, $p < .05$ (two-tailed).

Table 14.17. Correlations among P-E Fit Expectations Predictor Measures

Predictor	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
ABS Composite Scores																									
1. Supported Reinforcers																									
2. Unsupported Reinforcers	.02																								
3. Expects Recog/Achieve	**.45**	**.30**																							
PSES Scale Scores																									
4. Realistic	**.23**	.07	**.18**																						
5. Investigative	.12	**.34**	**.36**	**.38**																					
6. Artistic	.00	**.34**	**.23**	**.13**	**.52**																				
7. Social	**.18**	**.18**	**.19**	**.35**	**.34**	**.32**																			
8. Enterprising	**.37**	.00	**.26**	**.36**	**.22**	.12	**.41**																		
9. Conventional	**.40**	.08	**.18**	**.41**	.11	.01	**.51**	**.51**																	
AWKS Scale Scores	**.32**	.07	**.21**	**.20**	.06	.03	.07	**.29**	**.29**																
10. Achievement/Effort																									
11. Adaptability/Flexibility	**.38**	.09	**.28**	**.22**	.10	-.03	**.13**	**.22**	**.15**	**.27**															
12. Attention to Detail	**.35**	.05	**.23**	**.34**	.03	-.12	.12	**.23**	**.29**	**.38**	**.48**														
13. Concern for Others	.12	**.29**	**.14**	**.19**	**.27**	**.27**	**.20**	**.15**	.07	**.17**	**.30**	**.16**													
14. Cooperation	**.32**	.09	**.16**	**.18**	.05	-.09	**.13**	**.13**	**.26**	**.32**	**.52**	**.51**	**.23**												
15. Dependability	**.35**	.05	**.22**	**.25**	.07	-.09	**.21**	**.22**	**.26**	**.41**	**.49**	**.53**	**.24**	**.44**											
16. Energy	**.50**	.09	**.27**	**.30**	**.13**	-.03	**.22**	**.20**	**.26**	**.37**	**.45**	**.56**	**.29**	**.41**	**.52**										
17. Independence	.01	**.24**	.02	.02	**.15**	**.20**	.12	-.02	-.06	.10	**.17**	.07	**.32**	-.02	.05	.09									
18. Initiative	**.37**	**.17**	**.34**	**.25**	.09	.04	**.15**	**.20**	**.22**	**.37**	**.25**	**.42**	**.27**	**.22**	**.34**	**.38**	**.15**								
19. Innovation	.07	**.28**	**.22**	**.16**	**.29**	**.40**	**.24**	**.27**	.03	**.24**	**.27**	**.18**	**.38**	.11	.11	**.18**	**.31**	**.21**							
20. Leadership Orientation	**.25**	.11	**.25**	.12	.08	.08	**.16**	**.25**	**.26**	**.39**	**.36**	**.28**	**.17**	**.29**	**.28**	**.22**	**.17**	**.31**	**.35**						
21. Persistence	**.36**	.10	**.27**	**.29**	**.15**	.01	**.30**	**.36**	**.32**	**.39**	**.46**	**.56**	**.15**	**.39**	**.48**	**.47**	.10	**.44**	**.27**	**.34**					
22. Self-Control	**.30**	-.02	**.19**	**.22**	-.04	**-.15**	**.14**	**.26**	**.33**	**.43**	**.40**	**.58**	**.19**	**.45**	**.44**	**.41**	.04	**.48**	.10	**.37**	**.48**				
23. Social Orientation	**.44**	.02	**.25**	**.29**	.02	-.07	**.17**	**.22**	**.31**	**.39**	**.40**	**.54**	**.28**	**.50**	**.49**	**.45**	-.05	**.41**	**.21**	**.30**	**.49**	**.45**			
24. Stress Tolerance	**.42**	.02	**.24**	**.17**	.00	-.07	.12	**.17**	**.31**	**.45**	**.42**	**.53**	**.13**	**.38**	**.44**	**.49**	.06	**.39**	.06	**.31**	**.47**	**.57**	**.46**		
25. Cultural Tolerance	**.42**	.06	**.21**	**.27**	.04	.00	**.29**	**.27**	**.33**	**.38**	**.51**	**.46**	**.28**	**.50**	**.56**	**.49**	.08	**.45**	**.18**	**.34**	**.55**	**.48**	**.53**	**.49**	

Note. $n = 180$ to 324. Expects Recog/Achieve = Expects Army to Provide Recognition/Achievement. Bolded correlations are significant, $p < .05$ (two-tailed).

There were also numerous significant correlations between logically related PSES and AWKS scale scores. For example, PSES Realistic correlated .30 with AWKS Energy, PSES Artistic .40 with AWKS Innovation, and PSES Enterprising .36 with AWKS Persistent. At the same time, several unexpected relations emerged. AWKS Cooperation and Social Orientation scales, for example, were more related to PSES Conventional (*r* = .26 and .31) than to PSES Social (*r* = .13 and .17). Likewise, AWKS Leadership Orientation correlated as highly with PSES Conventional (*r* = .26) as it did with PSES Enterprising (*r* = .25). Nonetheless, the general pattern of results suggests that expectations scores were consistent across measures of similar values, interests, and personality constructs.

Table 14.18 displays correlations between expectations scores from the ABS and PSES and scores from the remaining predictors. For the ABS composite scores, the single largest correlation was .29 between ABS Expects Army to Provide Recognition/Achievement scores and WPS Investigative interests. The Supported Reinforcers composite was most related to PSJT Overall Judgment scores (*r* = .28) and several of the RBI scale scores (e.g., *r* = .26 with Hostility to Authority), whereas Unsupported Reinforcer scores were most related to IFQ scale scores (e.g., *r* = .28 with both Artistic and Social). As for the Recognition/Achievement composite, scores on it were positively correlated with most of the RBI scale scores (e.g., *r* = .28 with Army Affect), and all of the WPS scale scores (e.g., *r* = .29 with Investigative). Interestingly, ABS scores were more highly related to the personality and vocational interests measures than to work values composites of the WVI.[93]

As for the PSES, it is interesting that several of its scales were negatively related to cognitive ability. The negative correlations between ability and Investigative and Artistic scores could be due to the fact that recruits of lower ability have less accurate expectations about the Army work environment (i.e., because the Army environment does not support these two types of interests; see Chapter 13). At the same time, however, correlations between cognitive ability scores and expectations about interests the Army environment tends to support (e.g., Realistic, Social) were negative and/or small and statistically nonsignificant. Several of the PSES scale scores also related positively to PSJT Overall Judgment scores. As with cognitive ability, there were positive correlations between the PSJT and PSES interests the Army work environment is thought to support. In fact, the correlation of .33 between Judgment scores and Conventional interests was the strongest relationship between the PSES and the other predictors. Finally, PSES scores correlated significantly with several of the RBI scale scores, although many of the relationships did not appear to be theoretically meaningful (e.g., *r* = .31 between PSES Conventional and RBI Cultural Tolerance).

Correlations between AWKS scale scores and the other predictors are presented in Table 14.19. As with the other expectations measures, there was some evidence that cognitive ability was related to the accuracy of recruits' expectations about the Army work environment. For instance, AWKS Independence and Innovation scale scores were negatively related to AFQT and ASVAB scores, whereas Stress Tolerance was positively related cognitive ability. A similar pattern of relations was found with PSJT whereby higher Judgment scores tended to correlate

[93] This table contains correlations between needs and expectations scores associated with the same constructs (e.g., correlations between WVI and ABS scores). However, these relations are discussed in Chapter 13 and thus are not reiterated here.

positively with temperament variables supported by the Army environment (e.g., $r = .27$ with Dependability) and negatively related to temperament variables the Army does not necessarily support (e.g., $r = -.13$ with Independence).

Table 14.18. Correlations between P-E Fit Expectations Predictors (ABS and PSES Scale Scores) and the Other Predictor Measures

	ABS Composite Scores			PSES Scale Scores					
Predictor	Support Reinforce	Unsupport Reinforce	Recog/ Achieve	R	I	A	S	E	C
ASVAB Composite Scores									
AFQT	**.18**	**-.18**	.02	-.04	**-.18**	**-.25**	-.06	.05	**.14**
Verbal	.20	**-.13**	.11	.00	**-.15**	**-.27**	**-.13**	.08	.10
Quantitative	.05	**-.16**	-.06	-.06	**-.19**	**-.19**	-.06	.03	**.13**
Technical	**.18**	-.12	.13	-.02	-.06	**-.17**	**-.17**	.06	-.03
Spatial	.00	-.02	.10	-.01	-.08	-.05	-.03	.00	.06
Psychomotor Ability Test									
Time-to-Fire Scale	.00	**-.16**	.07	-.07	-.08	**-.14**	-.09	-.04	-.04
Precision Composite	-.04	-.14	-.02	.00	**-.15**	**-.13**	.08	.04	-.01
PSJT Overall Judgment	**.28**	**-.21**	.12	**.20**	.03	-.09	**.23**	**.28**	**.33**
Education Tier	.10	-.02	.04	-.03	.03	.01	.07	.08	.06
REPETE Scale Scores									
Computer Courses Taken	-.04	-.04	.01	-.09	-.08	-.04	-.08	-.07	**-.13**
Mean Level of Mastery	.08	.08	.12	.02	-.09	-.09	.01	.03	-.01
General Computer Skills	.04	.02	.12	.00	-.09	-.07	-.01	.02	-.04
Basic Computer Certs.	-.05	.14	.01	.03	.09	.02	-.01	-.03	-.05
Advanced Computer Certs.	-.01	**.18**	-.04	-.02	.00	-.01	.02	.06	-.07
RBI Scale Scores									
Peer Leadership	.10	-.01	**.16**	.07	.06	.01	**.14**	.10	.11
Cognitive Flexibility	**.17**	-.01	.08	.06	-.03	-.01	.09	**.12**	**.13**
Achievement Orientation	.10	-.06	.12	.11	.08	.07	**.20**	.09	**.14**
Fitness Motivation	.09	-.06	.12	**.13**	.10	.08	.10	.01	-.01
Diplomacy	.05	-.05	.12	.01	.00	.04	**.16**	.06	.11
Stress Tolerance	.11	-.01	**.18**	.03	.01	.02	.02	-.04	.08
Hostility to Authority	**-.26**	.09	-.14	**-.13**	-.11	-.05	**-.17**	**-.16**	**-.24**
Self-Esteem	**.23**	-.08	**.16**	**.23**	**.12**	.08	**.20**	**.20**	**.19**
Narcissism	.11	-.07	-.05	.10	.10	.10	**.16**	.10	.05
Cultural Tolerance	**.24**	-.01	**.19**	**.14**	.03	.04	**.25**	**.18**	**.31**
Internal Locus of Control	**.25**	-.11	.08	.06	.06	.01	.07	.07	**.14**
Army Identification	**.18**	-.06	**.28**	**.26**	**.27**	**.22**	**.23**	.07	**.16**
Respect for Authority	.12	.05	**.16**	**.12**	.00	-.01	**.15**	.08	**.19**
Lie Scale	.12	-.05	.12	**.14**	**.14**	.04	.08	.10	**.15**
WSI Scale Scores									
Achievement/Effort	.00	.12	.10	.01	-.03	-.01	.00	.02	.11
Adaptability/Flexibility	.10	.07	.08	.07	.06	.05	-.02	.04	.01
Attention to Detail	.06	-.13	**.16**	.03	.06	-.03	-.07	.03	-.02

Table 14.18. (continued).

Predictor	ABS Composite Scores			PSES Scale Scores					
	Support Reinforce	Unsupport Reinforce	Recog/ Achieve	R	I	A	S	E	C
WSI Scale Scores (cont.)									
Concern for Others	.11	.00	.02	-.05	.02	-.01	-.01	.01	.02
Cooperation	-.10	-.03	-.10	-.09	.01	.07	.00	-.06	-.06
Dependability	.05	-.01	.06	-.04	.08	.10	-.09	-.03	-.03
Energy	.08	.03	-.05	.01	-.04	-.01	-.06	-.05	-.10
Independence	.05	.00	-.03	-.07	-.04	-.07	-.03	.04	-.02
Initiative	.08	-.02	.05	**.15**	.05	-.01	.02	.08	.07
Innovation	-.09	.09	-.02	-.08	-.09	-.07	-.03	-.04	**-.13**
Leadership Orient	-.06	-.09	-.04	-.08	-.11	-.06	.04	-.04	-.04
Persistence	.00	-.07	-.02	.01	-.05	.01	.06	.07	.07
Self-Control	-.09	-.08	-.05	.10	.07	.00	.08	-.02	.03
Social Orientation	-.11	.07	-.05	.02	.05	**.13**	.07	-.02	.02
Stress Tolerance	-.03	.02	-.03	.06	.07	.00	-.01	-.05	.01
Cultural Tolerance	-.03	.02	-.04	-.02	-.09	-.04	.06	.02	.08
WVI Composite Scores									
Growth	.15	-.15	.12	**.18**	.12	-.01	.09	.13	**.13**
Status	.00	.00	.09	.01	**.15**	-.03	.10	.01	.05
Stimulation	.05	.02	.03	**.13**	.07	-.06	.02	.09	-.07
Comfort	-.04	.01	-.09	.00	.03	-.06	.07	.00	-.06
Altruism	.11	.09	**.19**	.11	.12	-.01	**.14**	.12	**.17**
Self-Direction	.07	-.04	-.07	-.06	-.01	-.08	-.02	.06	-.03
WPS Scale Scores									
Realistic	.01	.09	**.20**	**.15**	**.22**	**.13**	-.05	.02	-.09
Investigative	**.21**	.02	**.29**	.06	.07	-.02	.05	**.17**	**.13**
Artistic	.07	**.21**	**.16**	-.05	.06	.06	.06	.04	-.02
Social	**.14**	.08	**.18**	.08	**.11**	.07	**.15**	**.12**	**.14**
Enterprising	**.21**	.03	**.25**	.08	**.13**	.02	**.13**	**.12**	**.14**
Conventional	.07	.08	**.18**	.04	**.14**	**.15**	**.11**	.04	.09
IFQ Scale Scores									
Realistic	.00	.13	.05	.05	.10	.08	-.03	.04	-.02
Investigative	**.22**	**.20**	.11	**.14**	.07	-.01	**.18**	**.21**	**.22**
Artistic	.13	**.24**	.12	.07	.05	.03	.08	.07	.09
Social	.09	**.24**	.04	.06	.07	.02	**.14**	.08	**.13**
Enterprising	-.01	.12	.04	.06	.06	.05	**.11**	-.01	.00
Conventional	-.01	.13	.02	-.03	.06	**.11**	.10	.07	.10

Note. n = 147 to 315. Support Reinforce = Supported Reinforcers. Unsupport Reinforce = Unsupported Reinforcers. Recog/Achieve = Expects Army to Provide Recognition/Achievement. R = Realistic. I = Investigative. A = Artistic. S = Social. E = Enterprising. C = Conventional. Bolded correlations are significant, $p < .05$ (two-tailed).

Table 14.19. Correlations between P-E Fit Expectations Predictors (AWKS Scale Scores) and the Other Predictor Measures

	Army Work Knowledge Survey (AWKS) Scale Scores															
Predictor	Achieve/ Effort	Adapt/ Flexible	Attend Details	Concern Others	Cooperate	Depend	Energy	Indepndt	Initiative	Innovate	Leader Orient	Persist	Self-Control	Social Orient	Stress Tolerance	Cultural Tolerance
ASVAB Composite Scores																
AFQT	-.01	-.07	.02	**-.26**	.07	.08	.00	**-.17**	.02	**-.29**	-.09	.00	.12	.03	**.16**	-.05
Verbal	.04	-.04	.03	**-.25**	.09	.02	.01	**-.16**	.01	**-.26**	-.07	.03	.12	.05	**.22**	-.01
Quantitative	-.09	-.07	-.01	**-.23**	.05	.10	-.03	**-.16**	-.06	**-.25**	-.12	-.04	.03	-.03	.03	-.07
Technical	.00	.08	-.02	**-.21**	.05	-.03	-.08	-.12	-.07	**-.17**	.03	.05	.08	-.02	**.13**	-.01
Spatial	.05	.00	.07	**-.14**	.09	.12	.02	-.03	.04	-.06	.08	.06	.09	-.03	.08	.02
Psychomotor Ability Test																
Time-to-Fire Scale	-.09	**-.13**	-.11	**-.28**	**-.18**	.00	**-.18**	-.11	-.08	**-.28**	-.04	-.09	.00	-.09	.02	-.12
Precision Composite	-.02	**-.16**	-.04	**-.25**	-.12	-.04	**-.16**	**-.18**	-.12	**-.14**	-.01	-.03	.00	-.07	-.08	-.13
PSJT Overall Judgment	**.15**	.11	**.27**	**-.14**	**.17**	**.27**	**.17**	**-.13**	.12	-.05	.10	**.23**	**.24**	**.25**	**.26**	**.16**
Education Tier	.04	.05	.00	**.14**	.01	.01	-.02	.10	-.03	.09	.08	.04	-.04	.03	-.07	.01
REPETE Scale Scores																
Computer Courses Taken	**-.14**	-.09	-.04	-.08	-.06	-.01	-.01	-.01	-.06	-.06	-.04	**-.16**	-.05	**-.14**	-.05	**-.15**
Mean Level of Mastery	-.07	.03	.04	-.04	.09	.12	.12	.04	.05	-.02	.03	.03	.01	.07	.03	.09
General Computer Skills	**-.14**	.00	.00	-.04	.03	.07	.06	.02	-.05	-.06	-.03	-.06	-.07	-.03	-.03	-.03
Basic Computer Certs.	.00	-.01	.03	.01	-.04	.04	.06	.04	-.05	.04	.07	-.06	-.10	.00	-.03	-.09
Advanced Computer Certs.	-.06	.07	.08	.07	.06	.05	.09	.00	-.06	.05	.07	.02	-.06	.05	-.06	.06
RBI Scale Scores																
Peer Leadership	-.01	-.05	-.04	**-.15**	-.04	-.10	-.01	-.03	.06	.00	-.03	.00	-.01	-.01	-.03	.03
Cognitive Flexibility	**.17**	-.05	.05	-.10	.07	.02	.06	-.04	.10	-.09	-.05	.07	.06	.05	.02	**.14**
Achievement Orientation	.02	-.09	.07	**-.16**	-.02	-.03	.07	-.12	.02	.03	.01	.11	.02	.11	.02	.00
Fitness Motivation	-.01	.04	.01	-.01	-.01	.02	.08	.11	.10	.02	.04	.04	.10	.05	.07	.06
Diplomacy	.04	.02	.01	-.11	**.15**	-.01	.04	-.06	.05	.10	.10	**.17**	.09	.08	.10	**.17**
Stress Tolerance	-.05	**-.18**	**-.13**	**-.19**	-.06	-.05	-.11	-.08	-.05	-.10	-.06	-.11	-.07	**-.14**	-.11	-.05
Hostility to Authority	-.11	-.05	**-.17**	.00	**-.18**	**-.21**	-.10	**.14**	-.11	.02	-.02	**-.21**	**-.14**	**-.18**	-.10	**-.14**
Self-Esteem	**.22**	.12	**.21**	-.04	.14	.10	**.18**	-.02	**.14**	.06	.05	**.19**	**.16**	**.21**	**.18**	**.19**
Narcissism	.13	.04	**.14**	.03	.01	.08	**.19**	-.01	.05	.04	-.03	.08	.06	.12	.06	.07
Cultural Tolerance	.13	.08	**.15**	-.04	**.18**	.07	**.15**	-.06	.13	.07	.04	**.22**	**.14**	**.17**	**.20**	**.32**
Internal Locus of Control	.11	.10	**.17**	-.04	.08	**.18**	.11	.06	.10	-.01	.08	.11	**.14**	.10	**.14**	**.13**
Army Identification	.08	-.10	-.03	.03	-.03	-.08	.11	.00	.00	.02	-.02	-.03	.00	.00	.01	.02
Respect for Authority	-.01	.03	**.14**	.08	.12	**.15**	.08	**-.15**	-.03	-.04	-.04	.10	.10	**.14**	.07	.13
Lie Scale	.09	.01	.13	.00	.05	.05	-.03	-.10	.06	.00	**.14**	.10	.11	.03	.04	.03
WSI Scale Scores																
Achievement/Effort	.08	.03	**.16**	-.02	-.01	-.02	.08	.02	.07	.03	.10	.05	.04	.06	.06	-.04
Adaptability/Flexibility	.02	.11	.02	.09	.00	-.03	.04	.12	.03	.07	-.05	.03	-.06	.04	.01	.09
Attention to Detail	-.02	-.02	.11	-.01	-.03	-.08	.04	-.09	-.02	.05	.00	.04	.05	-.08	**.16**	-.10

Table 14.19. (continued)

Predictor	Army Work Knowledge Survey (AWKS) Scale Scores															
	Achieve/ Effort	Adapt/ Flexible	Attend Details	Concern Others	Cooperate	Depend	Energy	Indepndt	Initiative	Innovate	Leader Orient	Persist	Self-Control	Social Orient	Stress Tolerance	Cultural Tolerance
WSI Scale Scores (cont.)																
Concern for Others	.06	.03	.06	**.20**	.06	.05	.03	.00	.02	-.07	-.05	.00	-.03	.04	.01	.03
Cooperation	-.05	-.04	-.10	.09	-.04	-.07	-.09	.08	-.07	-.03	.03	-.01	-.05	-.02	-.08	.03
Dependability	-.03	-.04	-.04	-.07	.00	.04	.02	-.06	.07	.06	.03	-.02	.08	-.02	.03	-.04
Energy	-.04	-.12	-.07	-.06	-.01	-.03	-.04	.04	-.08	-.05	-.01	-.09	-.06	.05	-.03	-.06
Independence	.06	.04	.07	-.03	.04	**.14**	-.03	.11	.01	-.05	.02	.08	.07	-.05	.07	-.04
Initiative	.07	**.16**	.04	.00	.03	.06	.01	.01	.04	-.01	.00	.05	.06	.00	.10	.04
Innovation	-.08	-.01	.04	.03	-.03	-.02	.06	.07	.00	.08	.00	.00	-.01	-.06	-.07	.12
Leadership Orient	-.08	-.08	-.04	-.03	-.07	.03	.00	-.02	.03	-.03	.02	-.03	.00	.02	-.08	-.03
Persistence	.03	.07	.10	-.08	.11	.11	-.01	-.11	.02	-.06	-.03	**.16**	.12	.11	**.14**	.05
Self-Control	.09	.08	.04	-.13	.07	.04	.02	-.01	.00	.04	.07	.06	.06	-.05	.04	-.01
Social Orientation	-.05	-.12	**-.20**	.08	-.05	-.05	-.09	**-.15**	**-.16**	.00	-.01	**-.17**	-.13	.03	**-.19**	-.01
Stress Tolerance	-.12	-.02	**-.17**	-.07	-.10	**-.17**	-.09	-.05	-.02	.06	.00	**-.16**	-.11	-.03	-.08	**-.14**
Cultural Tolerance	.06	-.03	-.02	.00	.03	.01	.04	.03	.04	-.06	-.11	.02	-.02	-.04	-.06	.11
WVI Composite Scores																
Growth	**.16**	.01	**.17**	-.06	.10	.12	.09	-.03	.07	.00	.10	.08	.09	.13	.13	.09
Status	.08	.03	-.04	-.03	.01	.02	-.04	.04	-.03	.02	.03	.02	-.03	-.02	-.02	.05
Stimulation	**.15**	.12	.03	.05	.14	.02	**.21**	.09	.02	.14	.11	.05	.04	.07	.05	.09
Comfort	-.03	.02	-.12	.05	-.06	.05	-.02	.05	-.02	-.04	-.05	.04	-.10	-.07	.02	.07
Altruism	**.19**	-.05	.00	.01	-.05	-.05	.06	-.03	.00	-.01	-.04	.02	-.04	.09	.10	.02
Self-Direction	.06	.13	-.05	.05	.11	.08	-.01	.08	-.04	.08	-.01	.01	.07	-.06	.08	.02
WPS Scale Scores																
Realistic	.10	.09	.06	.12	.00	-.08	.01	.04	**.16**	**.20**	.12	.07	.12	.11	.03	.06
Investigative	**.14**	**.17**	**.25**	.03	**.18**	**.14**	**.13**	-.03	**.19**	.08	.04	**.17**	**.16**	**.15**	.12	**.14**
Artistic	.05	.06	.09	.11	.10	.06	**.13**	-.01	.08	.10	-.01	**.13**	.00	.04	-.06	**.13**
Social	**.17**	.10	**.19**	**.17**	.11	.11	**.23**	-.02	**.21**	.11	.01	**.20**	.10	**.16**	.09	**.16**
Enterprising	.09	.03	**.13**	.00	.05	.06	**.13**	-.04	**.21**	.08	.06	**.17**	.08	**.14**	.05	.11
Conventional	.11	.12	**.17**	.12	.12	**.14**	.08	.00	**.15**	**.13**	.06	.11	.09	**.14**	**.15**	.06
IFQ Scale Scores																
Realistic	.01	.12	-.06	.12	-.06	-.04	-.04	.07	.06	.09	.09	.04	**.16**	.01	.03	.02
Investigative	**.20**	.13	**.13**	.07	**.14**	**.16**	.11	.04	.11	.05	.08	**.22**	**.22**	**.13**	**.17**	**.25**
Artistic	**.13**	.12	.06	**.18**	.09	.09	**.15**	.01	.05	.09	-.01	.10	**.20**	.12	.09	**.18**
Social	**.19**	.08	.06	.09	.11	.03	.09	.07	.09	.06	.02	**.15**	**.15**	**.15**	**.14**	**.15**
Enterprising	.08	.06	-.01	.07	.03	.03	.09	.11	.07	.13	.02	.10	.11	.05	.04	.12
Conventional	**.14**	.07	.08	**.13**	.10	**.15**	-.03	.05	.10	.11	**.15**	.07	.13	**.15**	.06	.08

Note. n = 184 to 239. Achieve/Effort = Achievement/Effort. Adapt/Flexible = Adaptability/Flexibility. Attend Details = Attend to Details. Concern Others = Concern for Others. Cooperate = Cooperation. Depend = Dependability. Indepndt = Independence. Innovate = Innovation. Leader Orient = Leadership Orientation. Persist = Persistence. Social Orient = Social Orientation. Bolded correlations are significant, $p < .05$ (two-tailed).

AWKS scores also demonstrated some significant relations to RBI scores. Interestingly, some of the largest correlations were between AWKS and RBI scales of similar constructs, such as the Cultural Tolerance scales ($r = .32$) and AWKS Achievement/Effort and RBI Self-Esteem ($r = .22$). This suggests that recruits' expectations about the Army may have been influenced by their temperament. Alternatively, these relations may simply reflect the fact that recruits joined the Army because they thought it would fit their personality. It is noteworthy, however; that other theoretically linked scales, such as AWKS Leadership Orientation and RBI Peer Leadership, were not highly related. Finally, some AWKS scale scores were significantly correlated with the vocational interest scales. As with the RBI, some of these relations were sensible (e.g., $r = .25$ between AWKS Attention to Detail and WPS Investigative), whereas other relations were not (e.g., $r = .25$ between AWKS Cultural Tolerance and IFQ Investigative).

P-E Fit Indices

The correlations in Table 14.15 show the extent to which P-E fit work values and vocational interests needs measures assess similar underlying constructs. However, scores on these scales will not serve as fit measures for the concurrent validation but rather as input for computing fit indices. As discussed in Chapter 13, the specific algorithms for computing fit scores have not been finalized. We did, however, compute D2 and r based fit indices, which are the closest we have to the fit indices expected to be used in the validation research.

Table 14.20 displays correlations among the various P-E fit index scores we computed.[94] The high correspondence between the two types of fit indices (D^2 and r) and between the current and future Army indices for a given measure is discussed in Chapter 13. Thus, we focus on relations between fit indices based on scores from different instruments.[95] Several findings are noteworthy. First, the same types of fit indices from different instruments were, in general, positively related (e.g., $r = .20$ between the WVI and WPS Current Army r indices). This suggests that a good fit on one set of constructs (e.g., work values) tended to mean a good fit on another set of constructs (e.g., personality). However, correlations among the various fit indices were rather modest. Excluding relations between the WPS and IFQ fit indices (discussed in Chapter 13), the strongest correlation was between the PSES Current Army r index and the AWKS Current Army rank-order correlation ($r = .39$). Given this, there is an opportunity for the P-E fit index scores to explain unique variance in the criteria of interest (e.g., attrition).

Correlations between the P-E fit indices and scores on the remaining predictor measures are provided in Table 14.21.[96] Cognitive ability and PSJT scores tended to correlate positively with scores on the expectations-reality fit measures (i.e., the ABS, PSES, and AWKS). In fact, the correlation of .43 between PSJT Overall Judgment scores and the AWKS r fit index represents that strongest relationship between the fit indices and the other predictors. This suggests that recruits with higher ability and better judgment tended to have more accurate

[94] See Chapter 13 for more details about these fit indices.
[95] Recall from Chapter 13 that large rs and small D^2 values suggest a good fit between recruits' values/interests/temperament and those supported by the Army work environment. As such, negative correlations between the two indices indicate a similar relationship.
[96] This table does not include the work values, vocational interests, and temperament scales on which the P-E fit index scores are based.

Table 14.20. Correlations among P-E Fit Index Score Predictors

Fit Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
WVI																		
1. Current Army D^2																		
2. Current Army r	**-.56**																	
3. Future Army D^2	**.95**	**-.63**																
4. Future Army r	**-.53**	**.95**	**-.66**															
ABS																		
5. Current Army D^2	.00	-.08	.08	-.13														
6. Current Army r	.06	.03	-.03	.10	**-.77**													
WPS																		
7. Current Army D^2	.07	**-.22**	**.10**	**-.21**	**.15**	-.01												
8. Current Army r	**-.09**	**.20**	**-.13**	**.21**	.01	-.01	**-.33**											
9. Future Army D^2	**.09**	**-.25**	**.12**	**-.24**	**.15**	-.01	**.97**	**-.48**										
10. Future Army r	-.09	**.21**	**-.12**	**.21**	.03	-.01	**-.33**	**.98**	**-.50**									
IFQ																		
11. Current Army D^2	-.01	-.04	.00	-.03	.06	.08	**.35**	**-.11**	**.35**	**-.12**								
12. Current Army r	-.02	.04	-.01	.03	.06	-.06	**-.17**	**.51**	**-.25**	**.49**	**-.32**							
13. Future Army D^2	.00	-.06	.00	-.04	.05	.07	**.35**	**-.21**	**.38**	**-.22**	**.97**	**-.46**						
14. Future Army r	-.02	.06	-.01	.04	.05	-.04	**-.17**	**.51**	**-.26**	**.51**	**-.30**	**.96**	**-.48**					
PSES																		
15. Current Army D^2	-.08	.09	-.06	.05	.03	-.05	.02	.00	.00	.00	-.05	.03	-.05	.03				
16. Current Army r	.03	.02	-.03	.06	-.14	**.18**	-.09	-.03	-.08	-.02	-.03	**-.14**	-.01	**-.15**	**-.39**			
17. WSI Current Army r	**-.19**	**.31**	**-.22**	**.31**	-.08	.07	**-.13**	**.32**	**-.18**	**.29**	.01	**.13**	-.01	**.12**	.03	-.08		
18. AWKS Current Army r	-.03	.02	-.09	.08	**-.26**	**.35**	-.06	-.13	-.03	**-.14**	-.05	**-.14**	.00	**-.17**	-.03	**.39**	-.03	

Note. n = 146 to 661. WSI and AWKS Current Army *r*s are Spearman rank-order correlations. Bolded correlations are significant, $p < .05$ (two-tailed).

Table 14.21. Correlations between P-E Fit Index Score Predictors and the Other Predictor Measures

	WVI Fit Indices				ABS		WPS				IFQ				PSES			
Predictor	C D^2	C r	F D^2	F r	C D^2	C r	C D^2	C r	F D^2	F r	C D^2	C r	F D^2	F r	C D^2	C r	WSI r	AWKS r
ASVAB Composite Scores																		
AFQT	.04	-.05	.02	-.02	**-.28**	**.32**	.01	**-.16**	.05	**-.18**	.06	**-.11**	.07	**-.12**	-.09	**.26**	.00	**.36**
Verbal	.03	-.04	.03	-.03	**-.31**	.35	**.09**	**-.18**	**.14**	**-.20**	**.16**	**-.26**	**.19**	**-.24**	.00	**.22**	-.02	**.32**
Quantitative	.00	-.07	.00	-.05	-.11	.12	-.01	**-.13**	.03	**-.15**	.02	-.05	.04	**-.09**	-.02	**.16**	-.02	**.30**
Technical	.03	-.03	.04	-.02	**-.22**	**.22**	.02	-.05	.08	**-.13**	**.14**	-.03	**.14**	-.04	.05	.08	-.01	**.19**
Spatial	-.03	.03	-.02	.03	.00	-.03	-.04	-.07	.00	**-.10**	.08	-.05	**.10**	**-.08**	.02	.09	.02	.10
Psychomotor Ability Test																		
Time-to-Fire Scale	.02	-.04	.02	-.04	-.14	**.21**	.05	-.06	.08	**-.08**	**.15**	-.07	**.16**	-.07	.07	.02	.01	**.15**
Precision Composite	.01	.08	-.02	**.09**	-.03	.08	.00	-.04	.03	**-.08**	**.12**	**-.08**	**.13**	**-.08**	.07	.05	.08	**.14**
PSJT Overall Judgment	-.08	.07	**-.13**	**.10**	**-.25**	**.32**	**-.15**	-.01	**-.15**	.00	**-.14**	-.07	**-.10**	**-.10**	-.08	**.38**	.02	**.43**
Education Tier	-.03	.03	-.03	.03	-.01	.02	.06	.01	.07	-.01	.00	.02	.01	.00	.02	.01	-.02	-.13
REPETE Scale Scores																		
Computer Courses Taken	.00	.05	.01	.02	.00	.02	.00	-.04	.00	-.04	-.02	.00	-.01	-.02	-.04	-.10	.07	-.04
Mean Level of Mastery	-.02	.02	-.01	-.01	.00	.00	.03	**-.18**	.06	**-.17**	-.07	**-.10**	-.05	**-.09**	.07	.04	-.02	.08
General Computer Skills	.00	.04	.00	.02	.03	.02	.02	**-.18**	.05	**-.18**	**-.09**	**-.10**	-.05	**-.11**	.04	.03	-.02	.04
Basic Computer Certs.	.00	-.03	.00	-.03	.10	-.10	**.10**	**-.11**	**.12**	**-.11**	-.02	-.04	-.01	-.04	-.05	.01	.03	.00
Advanced Computer Certs.	-.02	-.01	-.02	.00	**.15**	**-.16**	**.14**	**-.11**	**.16**	**-.11**	.00	-.02	.02	-.03	-.04	.02	.00	.02
RBI Scale Scores																		
Peer Leadership	.05	.04	.03	.04	-.09	.10	.03	**-.18**	.03	**-.13**	-.05	**-.14**	-.05	-.06	**.13**	.03	**-.09**	.03
Cognitive Flexibility	-.01	**.10**	-.06	**.12**	-.09	.09	-.04	**-.34**	.02	**-.33**	**-.14**	**-.33**	-.05	**-.34**	**.14**	.07	**-.10**	**.16**
Achievement Orientation	**-.11**	**.19**	**-.14**	**.19**	-.06	.06	-.07	.00	**-.11**	.03	**-.13**	.00	**-.13**	.02	**.12**	.04	.08	.10
Fitness Motivation	-.06	**.18**	-.09	**.18**	-.11	**.15**	-.04	**.17**	-.06	**.13**	-.03	.04	-.05	.07	.06	.00	**.17**	-.03
Diplomacy	-.01	**.15**	-.03	**.14**	-.10	.12	-.01	-.07	-.04	-.02	-.02	-.05	-.03	.01	.07	.04	-.03	.02
Stress Tolerance	.00	**.16**	-.03	**.19**	**-.24**	**.19**	.00	.00	.02	-.02	.03	.00	.05	-.02	.08	-.05	**.12**	.02
Hostility to Authority	.02	-.05	.05	-.06	**.18**	**-.17**	.04	.03	.04	.03	-.01	**.08**	-.05	**.14**	-.05	-.07	-.04	**-.22**
Self-Esteem	.00	**.11**	-.05	**.14**	-.03	.07	.00	-.05	-.02	-.05	-.06	-.07	-.05	-.04	**.22**	.07	.01	**.15**
Narcissism	.02	.01	-.01	.01	.11	-.04	-.03	-.05	-.06	-.02	-.06	-.06	-.07	-.02	.07	.01	-.08	.10
Cultural Tolerance	-.06	**.15**	**-.12**	**.18**	**-.19**	**.18**	-.05	**-.09**	-.04	-.07	-.06	**-.14**	-.02	**-.14**	.06	**.13**	.04	.12
Internal Locus of Control	-.03	**.10**	-.07	**.11**	-.09	**.18**	.03	-.05	.01	-.04	.01	-.06	.02	-.07	.09	.07	.04	**.17**
Army Identification	**-.14**	**.36**	**-.21**	**.39**	**-.16**	**.16**	**-.11**	**.18**	**-.15**	**.17**	.01	-.02	.00	.01	.02	-.01	**.17**	-.09
Respect for Authority	-.01	**.14**	-.05	**.14**	-.12	.09	-.04	-.03	-.06	-.01	**-.12**	-.06	**-.09**	-.07	.05	.05	.05	**.16**
Lie Scale	-.02	.02	-.02	.03	.06	-.05	.05	-.01	.03	-.01	.04	.03	.05	.00	**.14**	-.04	-.03	.00

Note. n = 178 to 615. C D^2 = Current Army D^2 fit index. C r = Current Army r fit index. F D^2 = Future Army D^2 fit index. F r = Future Army r fit index. WSI and AWKS Current Army rs are Spearman rank-order correlations.

expectations about the nature of the Army work environment. In contrast, cognitive ability had either no relationship, or in some cases a negative relationship, to scores on the needs-supplies fit measures (i.e., WVI, WPS, IFQ, and WSI). For example, ability scores tended to correlate negatively to WPS and IFQ r based fit indices (and thus positively to the D^2 indices for these instruments). These results indicate that higher ability recruits tended to have values/interests/temperament that the Army environment does not support. The differential relations between cognitive ability and these two types of fit (i.e., expectations-reality fit and needs-supplies fit) is not that surprising given the low correspondence between recruits' needs and expectations (see Chapter 13).

As for the other predictor constructs, fit index scores were not consistently related to psychomotor ability or to education, training, and experience measures. However, there were numerous statistically significant correlations between P-E fit and RBI scale scores. For instance, given the covariation between RBI Cognitive Flexibility and cognitive ability (discussed earlier), it is not surprising that scores on this RBI scale tended to relate negatively to the WPS and IFQ r based indices (e.g., $r = -.34$ and -.33 with the current Army rs). It is also interesting that the WVI r based indices, which were largely unrelated to the other predictors, correlated positively with several of the RBI scales. Perhaps the most notable relationship was that between the fit index scores and RBI Army Affect scale scores. Specifically, recruits whose needs and expectations were more in line with the Army work environment tended to have greater affect for the Army (e.g., $r = .36$ with WVI current r index). In one sense, these relationships provide criterion-related validity evidence for the P-E fit measures, which were designed to predict attrition and its altitudinal precursors, including affective commitment toward the Army (see Chapters 7 and 13).

Summary

Results of the predictor cross instrument analyses revealed several noteworthy findings. For one, there appears to be minimal overlap between existing operational selection measures (i.e., AFQT composite scores and education tier) and the predictors being developed in this project. Specifically, the highest correlation between AFQT scores and the experimental predictors was -.29 (with AWKS Innovation) and the highest correlation for education tier was only .18 (with RBI Army Affect). These modest relations leave open the possibility for the experimental measures to provide incremental validity beyond the operational measures for predicting the performance of potential recruits. Likewise, although there were numerous statistically significant correlations among the experimental predictor scores, the magnitude of these correlations was generally quite modest. In fact, the largest cross instrument correlation was .51 (RBI Cognitive Flexibility and IFQ Investigative). Thus, despite the similarity of constructs assessed by many of the predictors (e.g., certain work values and temperament dimensions), redundancy in measurement does not appear to be a major concern.

The present results also provide additional construct-related validity evidence (beyond that described in earlier chapters) for many of the experimental predictors, particularly the RBI and the P-E fit needs and expectations measures. Indeed, correlations between scales from different instruments designed to assess the same or highly similar constructs tended to be larger than correlations between scales from different instruments designed to assess different constructs. It is important to note, however, that a more complete assessment of convergent and

discrimant validity should also include correlations between scale scores of different constructs within the same instrument (D. T. Campbell & Fiske, 1959). Although the cross-instrument correlations reported in this chapter provide evidence for convergent validity, their magnitude may be consistently smaller than the within-instrument correlations reported in earlier chapters, and thereby fail to provide discriminant validity evidence for these measures.

Perhaps the most disappointing result, from a construct validity perspective, was that PSJT temperament scales (particularly those based on items from the final version of the instrument) did not appear to relate consistently to scores on other temperament scales designed to measure similar constructs. In addition, the two forms of the PSJT temperament scales often correlated differently with other predictor measures, which suggest that the two forms are measuring different constructs. However, we reiterate that the PSJT was designed to measure specific Select21 KSAs (e.g., Adapts to Changing Situations, Exhibits Self-Management) that although conceptually related to temperament, are not temperament constructs per se.

Finally, as with the criterion intercorrelations, it is important to note the potential effects of measurement method on the predictor correlations we report. For example, some of the largest correlations were between scores on self-report measures that use Likert-type ratings (e.g., the RBI and the WPS and IFQ). In contrast, correlations between self-report and archival measures such as education tier and REPETE tended to be much lower. Thus, it is likely that at least some of the observed covariation among predictors is due to common method (and/or common "source") variance.

CHAPTER 15: SUMMARY COMMENTS AND NEXT STEPS

Deirdre J. Knapp (HumRRO), Trueman R. Tremble (Army Research Institute), and
John P. Campbell (HumRRO)

A Summary of Progress to Date

This report has documented development of predictor and criterion measures applicable for first-term enlisted Soldiers. Most of these instruments, which are summarized in Tables 15.1 and 15.2, will be used in a concurrent validation scheduled for 2005. The measures of interest were identified and designed based on results of a future-oriented job analysis (Sager, Russell, R. C. Campbell, & Ford, 2005). The criterion measures include performance rating scales to be completed by supervisors and peers, job knowledge tests, a situational judgment test, administrative performance indicators, and job/organization-related attitudes (e.g., satisfaction, commitment). This is a comprehensive set of measures that should allow us to examine how the predictor measures relate to various aspects of performance and other organizational fit indexes. Scores on the Select21 criterion measures are intended to reflect how well Soldiers might be expected to perform under future Army conditions.

Table 15.1. Summary of Select21 Criterion Measures

Title	Description
Peer and supervisor ratings	
Current Observed Performance Rating Scales (COPRS)	Scales for rating Soldiers' current performance. There are Army-wide and MOS-specific versions.
Future Expected Performance Rating Scales (FX)	Scales for rating expected Soldier effectiveness under anticipated future conditions. There are Army-wide and MOS/cluster-specific versions.
Tests	
Job Knowledge Tests	Selected response item (mostly multiple-choice) tests covering important job tasks. There are Army-wide and MOS-specific versions.
Criterion Situational Judgment Test (CSJT)	Realistic job problem situation items with four response options per item. Soldiers rate the effectiveness of each option (i.e., potential action to take in the situation).
Soldier self-report	
Select21 Personnel File Form (S21-PFF)	Personnel file information, most of which is also used as a basis for promotion qualification decisions. Includes awards, military education, Army Physical Fitness Test score, weapons qualification, deviance indicators (e.g., Article 15s), and indicators of exceptional performance (e.g., accelerated advancement).
Army Life Survey	Attitudinal criteria, including satisfaction (with the Army in general, supervision, peers, work, promotions, pay and benefits), organizational commitment (affective, continuance, and normative), perceived fit with the Army and MOS, perceived stress, career intentions, and belief in Army values.
Future Army Life Survey	Soldiers' anticipated reactions given various future conditions. Provides three composite scores – Future Fit, Future Stress, and Future Continuance.
Archival	
Separation status	Army records will be periodically checked to determine if, when, and for what reasons Soldiers separated from the Army. These data will be collected on individuals who participated in predictor pilot tests, field tests, and faking research.

Table 15.2. Summary of Select21 Predictor Measures

Title	Description
Armed Services Vocational Aptitude Battery (ASVAB)	This is a battery of 10 tests – General Science, Arithmetic Reasoning, Word Knowledge, Paragraph Comprehension, Auto Information, Shop Information, Math Knowledge, Mechanical Comprehension, Electronics Information, and Assembling Objects (AO). All but the last (AO) are used operationally to inform enlisted personnel selection and classification decisions.
Rational Biodata Inventory (RBI)	A primarily rationally constructed biodata measure tapping multiple personality constructs. Some items were drawn from several existing Army tests, and others were developed for the present application.
Work Suitability Inventory (WSI)	Respondents rank order 16 statements that describe their preferred work requirements (e.g., work that requires leading, taking charge, and giving direction). The statements also reflect personality constructs.
Psychomotor Tests	A Target Tracking test and a Target Shoot test, both adapted from tests developed in Project A.
Predictor Situational Judgment Test (PSJT)	Civilian-based situational items reflective of problems commonly encountered within the first few months of Army service, with four response options per item. Respondents rate the effectiveness of each option (i.e., potential action to take in the situation). There is an overall judgment score and temperament-based subscores.
Record of Pre-Enlistment Training and Experience (REPETE)	Prototype measure that allows respondents to self-report their education, certifications, and skill level associated with several content areas, with particular emphasis of 10 computer skill areas.
Work Values Inventory (WVI)	Respondents prioritize work values by rank ordering various reinforcers (e.g., opportunity to learn new skills) based on how important the reinforcers are to their ideal job.
Work Preferences Survey (WPS)	A Likert-type measure to assess interest in various work activities that reflect Holland's 6-factor RIASEC model.
Interest Finder Questionnaire (IFQ)	An adaptation of the DMDC-developed Interest Finder, this is a Likert-type measure of interest in various work activities that reflect Holland's 6-factor RIASEC model. It will be replaced in the concurrent validation with an updated version of the Interest Finder which is very similar to the Select21 IFQ.
Army Beliefs Survey (ABS)	A knowledge measure in which respondents indicate the extent to which they believe the Army supports the work values-type reinforcers reflected in the WVI.
Pre-Service Expectations Survey (PSES)	A knowledge measure in which respondents indicate the extent to which they believe the Army addresses each of Holland's 6 interest dimensions. Content parallels the dimensions covered by the WPS and IFQ.
Army Work Knowledge Survey (AWKS)	A knowledge measure in which respondents indicate the extent to which they believe the Army offers certain types of work requirements. Content parallels the WSI.

The Select21 predictor measures are intended to increment the level of prediction offered by the ASVAB, thus they focus primarily on variables that are distinct from cognitive abilities. There are three instruments that measure temperament variables, either directly or indirectly. There are also multiple instruments that use interests and work preferences to forecast organizational and job fit. The psychomotor tests offer considerable potential for classification. We also developed the prototype Record of Pre-Enlistment Training and Experience (REPETE) to demonstrate the potential utility of giving credit for job skills developed prior to entry. Of particular note in the Select21 predictor measurement research is our application of varied and innovative ways to address the potential for response distortion in an operational setting. Another innovation is basing person-environment fit predictors (and criteria) on a particularly thorough integration of the research in this area.

Concurrent Validation Plan

As discussed in Chapter 2, difficulties obtaining troop support from Army units prompted us to revise the Select21 research plan to minimize the support that would be required. Specifically, we reduced the number of target MOS samples from six to two, reduced our sample size requirements for the remaining samples (Army-wide, 11B, and 31U), and simplified our data collection requirements (e.g., limiting administration time to no more than one day per participating Soldier, collecting only one supervisor rating per Soldier).

We will administer all the measures we originally envisioned to the Army-wide sample. Criterion testing time requirements for the MOS-specific samples, however, will force reduction in the predictor measures to fit within a one-day testing window. We favored keeping predictors for the MOS samples that show particular potential for classification (e.g., the psychomotor tests) and dropping those predictors that would likely be used only for selection (e.g., the Predictor Situational Judgment Test; PSJT). Administration time was another relevant factor. Based on these considerations, we dropped the PSJT from the administration plan for the MOS-specific samples.

The original research plan called for collecting complete criterion and predictor data from 50 Soldiers in 20 MOS for a total Army-wide sample size of 1,000. Current plans are to collect complete data on 750 Soldiers in the Army-wide sample. We simplified the Army-wide request by asking for a mix of MOS rather than targeting certain MOS. This means it is unlikely we will be able to conduct hierarchical linear modeling (HLM) analyses as originally planned. We plan to keep the original 300 Soldier minimum sample size for the MOS samples.

Attrition Analysis Plans

Although not documented in the present report, we have developed an attrition analysis database that includes scores from the predictor pilot test, faking research, and predictor field test data collections. We will periodically update this file with separation status information for active component Soldiers. (Data for reserve component Soldiers is less accessible and separation status variables are not well defined, so they are not included in the attrition database.) Approximate sample sizes in this database are 180-300 per measure for the predictor pilot test cohort, 117-147 per measure for the faking research cohort, and 454 for the field test cohort. We have developed features that allow for the rapid generation of reports from this database as

needed for reporting to Army stakeholders (e.g., the Select21 Army Steering Committee [ASC]) and others (e.g., the Scientific Review Panel).

Other Sources of Evaluation Data

The Select21 research team has considered additional strategies that could be used to generate data for the evaluation of the experimental predictor measures should sufficient additional resources become available. We would like to conduct a limited data collection, perhaps using college students, to collect test-retest data on certain predictor measures because the reliability of some of the predictors cannot be estimated with internal consistency indexes (e.g., the Work Values Inventory and the Work Suitability Inventory). It would also be useful to look at the stability of temperament scores (particularly in samples of young people) and psychomotor test scores over time.

Since the beginning of this research program, we have talked about the possibility of administering the Select21 predictors to "special" samples that would offer additional insight into their power to predict performance. For example, such samples might allow an examination of predictive validity (e.g., by administering the predictors prior to participation in training or high fidelity field exercises) and of validity for predicting performance in settings that are particularly futuristic (e.g., in a Stryker Brigade). Some special samples might require development or revision of suitable criterion measures.

More recently, we have also considered other samples (not special samples, per se) to augment the support available from active component units. For example, reserve component units have deployed considerably in recent years; testing Soldiers in post-deploying reserve units might provide a useful augmentation.

Beyond Select21

No matter how positive the results might be, the Select21 research will not be sufficient for moving the experimental predictors into operational mode. Indeed, considerable additional policy, research, and administrative issues will need to be addressed prior to implementation. Required activities include the following:

- Conduct a longitudinal validation and/or initial operational test and evaluation (IOT&E)
- Collect additional data to inform use of predictors for classification
- Develop alternate forms and equating procedures
- Inform Army leaders
- Coordinate with the Department of Defense (DoD) and the other services
- Integrate new tests into an accession/classification decision model

Research Requirements

While we are expecting that several of the Select21 predictors will yield significant validity estimates in a concurrent validation, a conservative approach would be to follow the concurrent validation with a longitudinal validation of the most promising predictors. This would

be followed by an IOT&E as is the custom with new ASVAB forms. In an IOT&E, tests are administered to applicants and may be used for selection and classification decisions. After enough data have been collected, the tests are re-evaluated prior to full acceptance for continued use. If the concurrent validation results are positive, however, it would be worth considering moving straight to an IOT&E model. This would assess performance of the new tests in an operational environment. The IOT&E could include subsequent collection of training and/or job performance data on Soldiers who took the experimental tests prior to entry. Such research would also be helpful for reevaluating ASVAB composites and how they are used for classification decisions.

Additional classification research, involving performance data on a much larger number of MOS, is particularly significant given that the Select21 research plan has been scaled back from six MOS to two. That said, gathering enough data to determine exactly how tests should be used for classification decisions is not solely an issue with the new Select21 predictor tests. The ASVAB Assembling Objects (AO) test is not used operationally because the Army does not have the performance criterion data required to specify exactly how the scores should be used. The Army's ASVAB-based Aptitude Area scores (which are combinations of the other nine ASVAB tests) were originally derived using MOS performance data produced by the Skill Qualification Test (SQT) program and its predecessors. The SQT program was discontinued more than a decade ago, however, and has not been replaced with a comparable program. Discontinuation of the SQT program and the ensuing loss of administrative job performance data on job skills have hampered assessment of the AO test. As demonstrated in Select21, this loss requires that research on new predictors bear additional costs for performance test development and administration. Reinstituting Army job skills testing, as is being considered as part of ARI's "PerformM21" project (Knapp & R.C. Campbell, 2005), would make it feasible to use archival data to determine classification uses of pre-enlistment tests. Another avenue for supporting classification applications may be to revisit the synthetic validity strategies examined some time ago using Project A data (e.g., Wise, Peterson, Hoffman, J.P. Campbell, & Arabian, 1991).

Prior to, or concurrently with, additional validation research, it will be necessary to (a) determine what predictors require alternate forms (to support test compromise issues and re-testing requirements), (b) design procedures for developing and equating alternate forms, and (c) develop alternate forms in accordance with those procedures. For some measures (e.g., the Work Suitability Inventory), it may not be possible (nor necessary) to develop alternate forms. For an instrument like the Predictor Situational Judgment Test (PSJT), there are no commonly accepted strategies for creating alternate forms, so this will require some additional innovation.

Policy Requirements

To date, ARI has used the Select21 ASC as the primary vehicle for informing Army leadership about the products of this research. Additional policy input from the Army G-1 will be required, however, to shepherd the tests through the channels required for implementation and to make decisions regarding how the tests will be administered and used. Assuming any new tests would be administered through the Military Entrance Processing Command (USMEPCOM), coordination with the Department of Defense (DoD) and the other military services will be required. This will not be an easy task. ARI is currently paving the way by seeking opportunities to inform DoD and joint service representatives about this research program.

There are many questions related to how new selection and classification tests would be used. For example, psychomotor tests could be administered post-enlistment only to individuals being considered for certain MOS. Most other Select21 predictors would likely be most useful for selection, but how would scores on these tests be integrated with current selection requirements (i.e., Armed Forces Qualification Test [AFQT] scores, physical standards, moral screening)? One model would be to add additional initial selection hurdles (e.g., minimum scores on one or more new predictors). If minimum scores are to be set, this becomes the basis for another research requirement. Another model would be to develop a composite entry score that includes more than one new predictor. In any case, it would be desirable to conceive of a decision model that made use of the full range of scores rather than establishing a pass/fail criterion. Without a batch processing system for applicants, however, it may not be possible to do this. The Army's experimental Enlisted Personnel Allocation System (EPAS; Lightfoot, Ramsberger, & Greenston, 2000) would support such a system, but it has not been adopted for operational use.

Closing Remarks

Although the Select21 research will not provide all the information needed for implementing changes to the Army selection and classification system to better meet the needs of the 21st century, it is a strong start. The concurrent validation will provide basic answers as to which experimental predictors hold promise for appreciably improving upon the selection and classification power of the ASVAB.

REFERENCES

Ackerman, P.L. (1986). Individual differences in information processing: An investigation of intellectual abilities and task performance during practice. *Intelligence, 10,* 101-109.

Ackerman, P.L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin, 102,* 3-27.

Ackerman, P.L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General, 117,* 288-318.

Ackerman, P.L. (1989). Individual differences and skill acquisition. In P.L. Ackerman, R.J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences: Advances in theory and research.* New York: Freeman.

Agresti, A. (2002). *Categorical data analysis, 2nd edition.* Hoboken, NJ: John Wiley & Sons, Inc.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50,* 179-211.

Barrick, M. R., Mount, M. K., & Gupta, R. (2003). Meta-analysis of the relationship between the five-factor model of personality and Holland's occupational types. *Personnel Psychology, 56,* 45-74.

Barrick, M.R., & Mount, M.K. (1996). Effects of impression management and self-deception on the predictive validity of personality constructs. *Journal of Applied Psychology, 81,* 261-272.

Borman, W.C., Kubisiak, U.C., & Schneider, R.J. (1999). Work styles. In N.G. Peterson, M.D. Mumford, W.C. Borman, P.R. Jeanneret, & E.A. Fleishman (Eds.), *An occupational information system for the 21st century: The development of O*NET* (pp. 213-226). Washington D.C.: American Psychological Association.

Bradlow, E. T., Hoch, S. J., & Hutchinson, J. W. (2002). An assessment of basic computer proficiency among active Internet users: Test construction, calibration, antecedents and consequences. *Journal of Educational and Behavioral Statistics, 27 (3),* 237-253.

Budescu, D.V. (1988). On the feasibility of multiple matching tests – Variations on the theme by Gulliksen. *Applied Psychological Measurement, 12,* 5-14.

Cable, D. M., & DeRue, S. D. (2002). The convergent and discriminant validity of subjective fit perceptions. *Journal of Applied Psychology, 87,* 875-884.

Cable, D. M., & Edwards, J. R. (2004). Complementary and supplementary fit: A theoretical and empirical integration. *Journal of Applied Psychology, 89*(5), 822-834.

Cable, D. M., & Judge, T. A. (1996). Person-organization fit, job choice decisions, and organizational entry. *Organizational Behavior and Human Decision Processes, 67,* 294-311.

Campbell, C. H., Ford, P., Rumsey, M. G., Pulakos, E. D., Borman, W. C., Felker, D. B., De Vera, M. V., & Riegelhaupt, B. J. (1990). Development of multiple job performance measures in a representative sample of jobs. *Personnel Psychology, 43*, 277-300.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Campbell, J.P., McCloy, R.A., Oppler, S.H., & Sager, C.E. (1993). A theory of performance. In N. Schmitt, W.C. Borman, & associates (Eds.) *Personnel selection in organizations* (pp. 35-70). San Francisco: Jossey-Bass Publishers.

Campbell, J.P., & Knapp, D.J. (Eds.) (2001). *Exploring the limits in personnel selection and classification.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Campbell, R.C., Keenan, P.A., Moriarty, K.O., Knapp, D.J., & Heffner T.S. (2004). Army enlisted personnel competency assessment program Phase I (Volume II): Demonstration Competency Assessment Program development report (Technical Report 1152). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Cattin, P. (1980). Estimation of the predictive power of a regression model. *Journal of Applied Psychology, 65*, 407-414.

Clifton, T.C., Kilcullen, R.N., Reiter-Palmon, R., & Mumford, M.D. (1992, August). *Comparing different background data scaling procedures using triple cross validation.* Paper presented at the 100th annual meeting of the American Psychological Association, Washington DC.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Cohen, R.A. (1999). *An introduction to PROC LOESS for Local Regression.* Paper 273 at the 24th SAS Users Group International Conference. Online: http://support.sas.com/rnd/app/papers/loesssugi.pdf.

Cronbach, L. J. (1958). Proposals leading to analytic treatment of social perception scores. In R. Tagiuri & L. Petrullo (Eds.), *Person perception and interpersonal behavior* (pp. 353-379). Stanford, CA: Stanford University Press.

Cronbach, L. J., & Gleser, G. C. (1953). Assessing similarity between profiles. *Psychological Bulletin, 50*, 456-473.

Cullen, M. J., Sackett, P. R., & Lievens, F. (2004, April). *Threats to the operational use of situational judgment tests in the college admission process.* Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Orlando, FL.

Dawis, R. V., England, G. W., & Lofquist, L. H. (1964). A theory of work adjustment. *Minnesota studies in vocational rehabilitation, XV.* Minneapolis: University of Minnesota.

Dawis, R. V., & Lofquist, L. H. (1984). *A psychological theory of work adjustment.* Minneapolis: University of Minnesota Press.

Department of Labor (1970). *General Aptitude Test Battery Manual, Section III.* Washington, DC: Employment and Training Administration, U.S. Employment Service.

Department of the Army (1998, October). *Physical Fitness Training (FM 21-20).* Washington D.C.: Headquarters, Department of the Army.

Department of the Army (2000, May). *Promotion point worksheet (DA Form 3355).* Washington D.C.: Headquarters, Department of the Army.

Department of the Army (2004, January). *Enlisted promotions and reductions (AR 600-8-19).* Washington D.C.: Headquarters, Department of the Army.

Department of the Army (2003, April). *Army training and education (AR 350-1).* Washington D.C.: Headquarters, Department of the Army.

Drewes, D.W. (2000). Beyond the Spearman-Brown: A structural approach to maximal reliability. *Psychological Methods, 5,* 214-227.

Dyer, J. L., & Martin, G. H. (1999). *The computer background of infantrymen: FY99* (Research Report 1751). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Edwards, J. R. (1991). Person-job fit: A conceptual integration, literature review and methodological critique. *International review of industrial/organizational psychology* (Vol. 6, pp. 283-357). London: Wiley.

Edwards, J. R. (1993). Problems with the use of profile similarity indices in the study of congruence in organizational research. *Personnel Psychology, 46,* 641-665.

Edwards, J. R. (1996). An examination of competing versions of the person-environment fit approach to stress. *Academy of Management Journal, 39,* 292-239.

Edwards, J. R., & Parry, M. E. (1993). On the use of polynomial regression equations as an alternative to difference scores in organizational research. *Academy of Management Journal, 36,* 1577-1613.

Feldt, L.S., & Brennan, R.L. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 105-146). New York: American Council on Education/Macmillan Publishing Company.

Fleishman, E.A. (1967). Performance assessment based on an empirically derived task taxonomy. *Human Factors, 9,* 349-366.

Ford, L. A., Campbell, R. C., Campbell, J. P., Knapp, D. J, & Walker, C. B. (2000). *21st Century Soldiers and noncommissioned officers: Critical predictors of performance* (Technical Report 1102). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Gay, E. G., Weiss, D. J., Hendel, D. D., Dawis, R. V., & Lofquist, L. H. (1971). *Minnesota studies in vocational rehabilitation: xxviii: Manual for the Minnesota Importance Questionnaire.* Minneapolis: University of Minnesota.

Grafton, F.C., Czarnolewski, M.Y., & Smith, E.P. (1988). *Relationship between Project A psychomotor and spatial tests and TOW gunnery performance: A preliminary investigation* (SCTA Working Paper RS-WP-87-100). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Griffeth, R. W., Hom, P. W., & Gaertner, S. (2000). A meta-analysis of antecedents and correlates of employee turnover: Update, moderator tests, and research implications for the next millennium. *Journal of Management, 26*, 463-488.

Heffner, T.S., & Gade, P.A. (2003). Commitment to nested collectives in special operations forces. *Military Psychology, 15*, 209-224.

Hesketh, B., & Gardner, D. (1993). Person-environment fit models: A reconceptualization and empirical test. *Journal of Vocational Behavior, 42*, 315-332.

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74*, 167-184.

Hoehn, A.J., Wilson, T.R., & Richards, J.A. (1972). *Recruits' civilian-acquired skills: their potential value and their utilization in initial military assignments* (AFHRL-TR-72-16). Brooks, Air Force Base: Air Force Human Resources Laboratory.

Holland, J. L. (1985). *Making vocational choices: A theory of vocational personalities and work environments* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Hom, P. W., & Griffeth, R. W. (1995). *Employee turnover.* Cincinnati, OH: South-Western.

Hom, P. W., Griffeth, R. W., Palich, L. E., & Bracker, J. S. (1999). Revisiting met expectations as a reason why realistic job previews work. *Personnel Psychology, 52*, 97-112.

Hough, L.M. (1996, April). *Personality measurement and personnel selection: Implementation issues.* Paper presented at the 11th annual meeting of the Society of Industrial and Organizational Psychology, San Diego, CA.

Hough, L.M. (1997, April). *Issues and evidence: Use of personality variables for predicting job performance.* Paper presented at the 12th annual meeting of the Society of Industrial and Organizational Psychology, St. Louis, MO.

Hough, L.M. (1998). Effects of intentional distortion in personality measurement and evaluation of suggested palliatives. *Human Performance, 11*, 209-244.

Hough, L.M., Eaton, N.K., Dunnette, M.D., Kamp, J.D., & McCloy, R.A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75*, 581-595.

Hosmer, D. W., & Lameshow, S. (2000). *Applied logistic regression, Second edition*. New York: John Wiley and Sons, Inc.

Humphreys, L.G., & Montanelli, R.G. (1975). An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research, 10*, 193-206.

Hunter, J.E., & Hunter, R.F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*(1), 72-98.

International Personality Item Pool (2001). *A scientific collaboratory for the development of advanced measures of personality traits and other individual differences*. [On-line]. Available: http://ipip.ori.org/.

Jackson, D.N., Wrobleski, V.R., & Ashton, M.C. (2000). The impact of faking on employment tests: Does forced-choice offer a solution? *Human Performance, 13*, 371-388.

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69*, 85-98.

Jensen, A.R. (1980). *Bias in mental testing*. New York: The Free Press.

Jex, S. M. (1998). *Stress and job performance: Theory, research and implications for managerial practice*. Thousand Oaks, CA: Sage Publications.

Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research, 35(1)*, 1-19.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Chicago: Scientific Software International.

Judd, C.M., Kenny, D.A., & McClelland, G.H. (2001). Estimating and testing mediation and moderation in within-subject designs. *Psychological Methods, 6,* 115-134.

Karrasch, A.I. (2003). Antecedents and Consequences of Organizational Commitment. *Military Psychology, 15,* 225-236.

Kilcullen, R.N., Goodwin, J, Chen, G., Wisecarver, M., & Sanders, M. (2002). *Identifying agile and versatile officers to serve in the objective force*. Paper presented at the 2002 Army Science Conference, Orlando, FL.

Kilcullen, R.N., Mael, F.A., Goodwin, G.F., & Zazanis, M.M. (1999). Predicting U.S. Army Special Forces field performance. *Human Performance in Extreme Environments, 4(1)*, 53-63.

Kilcullen, R.N., White, L.A., Mumford, M.D., & Mack, H. (1995). Assessing the construct validity of rational biodata scales. *Military Psychology, 7(1)*, 17-28.

Kilcullen, R.N., White, L.A., Sanders, M., & Hazlett, G. (2003). *The Assessment of Right Conduct (ARC) administrator's manual. Research Note 2003-09).* Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Kilcullen, R.N., White, L.A., Zaccaro, S.A., & Parker, C. (2000). *Predicting managerial and executive job performance.* Paper presented at the 15th annual meeting of the Society of Industrial and Organizational Psychology, New Orleans, LA.

Knapp, D. J. (Ed.) (2003). Select21 measure development progress report *(IR-03-74). Alexandria, VA: Human Resources Research Organization.*

Knapp, D. J., Burnfield, J. L., Sager, C. E., Waugh, G. W., Campbell, J. P., Reeve, C. L., Campbell, R. C., White, L. A., & Heffner, T. S. (2002). *Development of predictor and criterion measures for the NCO21 research program* (Technical Report 1128). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Knapp, D.J., & Campbell, R.C. (Eds.) (2005). *Army enlisted personnel competency assessment program: Phase II report* (FR 05-06). Alexandria, VA: Human Resources Research Organization.

Knapp, D.J., McCloy, R.A., & Heffner, T.S. (Eds.) (2004). *Validation of measures designed to maximize 21st-century Army NCO performance* (Technical Report 1145). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Knapp, D.J., Waters, B.K., & Heggestad, E.D. (Eds.) (2002). *Investigations related to the implementation of the Assessment of Individual Motivation (AIM)* (Study Note 2002-02). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Ko, J-W., Price, J. L., & Mueller, C. W (1997). Assessment of Meyer and Allen's three-component model of organizational commitment in South Korea. *Journal of Applied Psychology, 82*, 961-973.

Kristof, A. L. (1996). Person-organization fit: An integrative review of its conceptualizations, measurements, and implications. *Personnel Psychology, 49*, 1-49.

Lightfoot, M.A., Ramsberger, P.F., & Greenston, P.M. (2000). Matching recruits to jobs: The Enlisted Personnel Allocation System (EPAS) (Special Report 41). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Manual for the Administration of the FY04 Common Task Test. (2003). [On-line], Available: http://alphacompany.us/CTT%20Manual%20FY04.pdf.

Marsh, L. C., & Cormier, D. R. (2002). *Spline regression models.* Thousand Oaks, CA: Sage.

Maurer, T.J., Raju, N.S., & Collins, W.C. (1998). Peer and subordinate performance appraisal measurement equivalence. *Journal of Applied Psychology, 83*, 693-702.

McCloy, R.A., Putka, D.J., Van Iddekinge, C.H., & Kilcullen, R.N. (2003). Temperament measures. In D.J. Knapp (Ed.), *Select21 measure development progress report* (IR-03-74) (pp. 30-39). Alexandria, VA: Human Resources Research Organization.

McCloy, R. A., Waugh, G., Medsker, G., Wall, J., Rivkin, D., & Lewis, P. (1999). *Development of the O*NET computerized work importance profiler*. Raleigh, NC: National Center for O*NET Development.

McCoy, S., Carr, C.L., Marks, P.V., & Mbarika, V. (2004). *Electronic versus paper surveys: Analysis of potential psychometric biases.* Paper presented at the annual meeting of the International Conference on System Sciences, Honolulu.

McDaniel, M.A., Morgeson, F.P., Finnegan, E.B., Campion, M.A., & Braverman, E.P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*, 730-740.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30-46.

McHenry, J.J. & Rose, S.R. (1986). *The validity and potential usefulness of psychomotor ability tests for personnel selection and classification* (Research Note 88-13). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Meyer, J. P., & Allen, N. J. (1991). The measurement and antecedent of affective, continuance and normative commitment to the organization. *Journal of Occupational Psychology, 63*, 1-18.

Meyer, J. P., & Allen, N. J. (1997). *Commitment in the workplace: Theory, research, and application.* Thousand Oaks, CA: Sage Publications.

Meyer, J. P., Allen, N. J., & Smith, C. A. (1993). Commitment to organizations and occupations: Extension and test of a three-component model. *Journal of Applied Psychology, 78*, 538-551.

Motowidlo, S. J., Diesch, A. C., & Jackson, H. L. (2003, April). *Relations between traits and ratings of behavioral effectiveness in situational judgment items*. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Orlando, FL.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75*, 640-647.

Mueller-Hanson, R., Heggestad, E. D., & Thornton, G. C., III (2003). Faking and selection: Considering the use of personality from a select-in and a select-out perspective. *Journal of Applied Psychology, 88*, 348-355.

Mumford, M.D., Uhlman, C.E., & Kilcullen, R.N. (1992). The structure of life history: Implications for the construct validity of background data scales. *Human Performance, 5*, 109-137.

Mumford, M.D., & Owens, W.A. (1987). Methodology review: Principles, procedures, and findings in the application of background data measures. *Applied Psychological Measurement, 2*, 1-31.

Mumford, M.D., & Stokes, G.S. (1991). Developmental determinants of individual action: Theory and practice in the application of background data. In M.D. Dunnette (Ed.) *The handbook of industrial and organizational psychology* (2nd edition). Orlando, FL.: Consulting Psychologists Press.

Muthen, L.K., & Muthen, B. O. (2001) *Mplus user's guide, 2nd edition*. Los Angeles, CA: Muthen & Muthen.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (Third ed.). New York: McGraw-Hill, Inc.

Ones, D. S., Viswesvaran, C., & Korbin, W. P. (1995, May). *Meta-analyses of fakability estimates: Between-subjects versus within-subjects designs*. Paper presented at the 10th annual meeting of the Society of Industrial and Organizational Psychology, Orlando, FL.

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology, 81*, 660-679.

Owens, W.A. (1976). Background data. In M.D. Dunnette (Ed.), *Handbook of industrial psychology*. Chicago: Rand-McNally.

Peterson, N.G. (Ed.) (1987). *Development and field test of the Trial Battery for Project A* (Technical Report 739). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Peterson, N.G., Russell, T.L., Hallam, G. Hough, L., Owens-Kurtz, C., Gialluca, K., & Kerwin, K. (1990). Analysis of the experimental predictor battery. In. J.P. Campbell & L.M. Zook (Eds.) *Building and retaining the career force: New procedures for accessing and assigning Army enlisted personnel: Annual Report 1990 FY* (Research Note 952). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Peterson, N. G., Smith, D., Hoffman, R. G., Pulakos, E. D., Reynolds, D., Potts, B. C., Oppler, S. H., & Whetzel, D. L. (1993). *Expanding the Concept of Quality in Personnel: Base period final report*. Washington, D.C.: American Institutes for Research.

Putka, D. J. (2004, September). *Plans for the Select21 attrition database and attrition-related analyses* (Memorandum for Record). Alexandria, VA: Human Resources Research Organization.

Putka, D.J., (2005, April). Using spline regression models to examine person-environment fit. In C. Sager (Chair), *Assessing person-environment fit for selection*. Symposium conducted at the 20th Annual Society for Industrial and Organizational Conference, Los Angeles, CA.

Putka, D. J., & Campbell, R. C. (2004). Simulated promotion point worksheet (SimPPW). In D. Knapp, R. McCloy, & T. Heffner (Eds.), *Validation of Measures Designed to Maximize 21st-Century Army NCO Performance* (Technical Report 1145). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Putka, D. J., & Strickland, W. J. (2004). *A comparison of the FY03 and FY99 first term attrition study cohorts* (DR-04-35). Alexandria, VA: Human Resources Research Organization.

Ramsberger, P. F., Wetzel, E. S., Sipes, D. E., & Tiggle, R. B. (1999). *An assessment of the values of new recruits* (FR-WATSD-99-16). Alexandria, VA: Human Resources Research Organization.

Reilly, R.R., & Chao, G.T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology, 35,* 1-62.

Riegelhaupt, B. J., Harris, C. D., & Sadacca, R. (1987). *The development of administrative measures as indicators of Soldier effectiveness.* (Technical Report 754). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Rosse, J.G., Stecher, M.D., Miller, J.L., & Levin, R. (1998). The impact of response distortion on pre-employment personality testing and hiring decisions. *Journal of Applied Psychology, 83,* 634-644.

Russell, T.L. & Katkowski, D. (2004) *Pilot test results.* Memorandum for the record. Alexandria, VA: Human Resources Research Organization.

Russell, T.L, Reynolds, D.H., & Campbell, J.P. (1994). *Building a joint-service classification research roadmap: Individual differences measurement* (AL/HR-TP-1994-0009). Brooks Air Force Base, TX: Armstrong Laboratory.

Sackett, P. R., & Mavor, A. (Eds.) (2002). *Attitudes, aptitudes, and aspirations of American youth: Implications for military recruitment.* Washington, D.C: National Academies Press.

Sager, C.E., Peterson, N.G., Oppler, S.H., Rosse, R.L., & Walker, C.B. (1997). An examination of five indexes of test battery performance: Analysis of the ECAT Battery. *Military Psychology, 9(1),* 97-120.

Sager, C. E., Russell, T. L., Campbell, R. C., & Ford, L. A. (2005). *Future Soldiers: Analysis of entry-level performance requirements and their predictors* (IR 03-82). Alexandria, VA: Human Resources Research Organization.

Saks, A. M., & Ashforth, B. E. (1997). A longitudinal investigation of the relationships between job information sources, applicant perceptions of fit, and work outcomes. *Personnel Psychology, 50,* 395-426.

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual difference constructs. *Psychological Methods, 8,* 206-224.

Schmitt, N., & Chan, D. (1998). *Personnel selection: A theoretical approach.* Thousand Oaks, CA: Sage.

Schmitt, N., Gooding, R. Z., Noe, R. D., & Kirsch, M. (1984). Meta-analysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37*, 407-422.

Schneider, B. (1987). The people make the place. *Personnel Psychology, 40*, 437-453.

Schoenfeldt, L.F. (1989). *Biographical data as the new frontier in employee selection research.* Address presented at the annual meeting of the Division of Evaluation, Measurement and Statistics of the American Psychological Association, New Orleans.

Schwab, D.P., & Oliver, R.L. (1974). Predicting tenure with biographical data: Exhuming buried evidence. *Personnel Psychology, 27*, 125-128.

Schwartz, S. H. (1994). Are there universal aspects in the structure and contents of human values? *Journal of Social Issues, 50*, 19-45.

Sherman, J. (1967). Problem of sex differences in space perception and aspects of intellectual functioning. *Psychological Review, 74,* 290-299.

Smith, E.P., & Graham, S.E. (1987). *Validation of psychomotor and perceptual predictors of Armor Officer M-1 gunnery performance* (Technical Report 766). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Smith, E.P., & Walker, M.R. (1987) Testing psychomotor and spatial abilities to improve selection of TOW gunners. *Proceedings of the Military Testing Association Meeting.*

Spector, P. E. (1997). *Job satisfaction: Application, assessment, cause, and consequences.* Thousand Oaks, CA: Sage Publications.

Strickland, W. J. (2004) (Ed.). *A longitudinal study of first term attrition and reenlistment among FY1999 enlisted accessions* (DR-04-14). Alexandria, VA: Human Resources Research Organization.

Teplitsky, M. L. (1991). *Junior Army officer retention intentions: A path analytic model* (Technical Report 934). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Tinsley, H. E. A. (2000). The congruence myth: An analysis of the efficacy of the person-environment fit model. *Journal of Vocational Behavior, 56*, 147-179.

Tomsic, M.L., Hendel, D.D., & Matross, R.P. (2000). *A World Wide Web response to student satisfaction surveys: Comparisons using paper and Internet formats.* Paper presented at the annual meeting of the Association of Institutional Research, Cincinnati, OH.

Toquam, J., Peterson, N., Rosse, R., Ashworth, S., Hanson, M., & Hallam, G. (March, 1986). *Concurrent validity data analyses: Cognitive paper-and pencil and computer-administered predictors.* A paper presented at the Project A IPR & SAG Meeting in Alexandria, VA.

Tremble, T.R., Payne, S.C., Finch, J.F., & Bullis, R.C. (2003). Opening organizational archives to research: Analog measures of organizational commitment. *Military Psychology, 15*, 167-190.

Uhlman, C.E., Reiter-Palmon, R., & Connelly, M.S. (1990). *A comparison and integration of empirical keying and rational scaling of biographical data items*. Paper presented at the Southeastern Psychological Association.

U.S. Army. (2001). *Concepts for the Objective Force, United States Army white paper*. Online at: http://www.army.mil/features/WhitePaper/default.htm.

U.S. Army. (2002, December). *Objective Force in 2015 white paper*. Arlington, VA: Department of the Army: Objective Force Task Force.Verquer, M. L., Beehr, T. A., & Wagner, S. H. (2003). A meta-analysis of relations between person-organization fit and work attitudes. *Journal of Vocational Behavior, 63*, 673-689.

Wainer, D., & Thissen, H. (2001). True score theory: The traditional method. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 44-47). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Walker, C.B. (1985). *The fakability of the Army's Military Applicant Profile (MAP)*. Paper presented at the Association of Human Resources Management and Organizational Behavior proceedings, Denver, CO.

Wall, J. E., Wise, L. L., & Baker, H. E. (1996). Development of the Interest-Finder – A new RIASEC-based interest inventory. *Measurement and Evaluation in Counseling and Development, 72*, 134-152.

Wanous, J. P. (1992). *Organizational entry* (2nd Ed.), Reading, MA: Addison-Wesley.

White, L.A., Gregory, E.L., Kilcullen, R.N., Galloway, E., & Nedegaard, R. (2001). *New maturity assessment tools for U.S. Army correctional specialists*. Paper presented at the 109th meeting of the American Psychological Association, San Francisco, CA.

White, L.A., & Kilcullen, R.N. (1992). *The validity of rational biodata scales*. Paper presented at the 100th meeting of the American Psychological Association, Washington DC.

White, L. A., & Young, M. C. (1998, August). *Development and validation of the Assessment of Individual Motivation (AIM)*. Paper presented at the Annual Meeting of the American Psychological Association, San Francisco, CA.

Wise, L.L., Peterson, N.G., Hoffman, R.G., Campbell, J.P., & Arabian, J. (Eds.) (1991). *Army synthetic validity project: Report of Phase III results* (Vol. I) (Technical Report 922). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Wright, S. S., & Miederhoff, P. A. (1999). Selecting students with personal characteristics relevant to pharmaceutical care. *American Journal of Pharmaceutical Education, 63*, 132-138.

Zickar, M. J. (2000). Modeling faking on personality tests. In D. Ilgen & C. L. Hulin (Eds.), *Computational modeling of behavioral processes in organizations* (pp. 95-108). Washington, DC: American Psychological Association.

APPENDIX A

ARMY-WIDE PERFORMANCE DIMENSIONS

1. *Performs Common Tasks.* Possesses the necessary knowledge and skill to perform common tasks at the appropriate skill level (e.g., land navigation, field survival techniques, and chemical, biological, radiological and nuclear [CBRN] protection).

2. *Solves Problems and Makes Decisions.* Reacts to new problem situations by applying previous experience and previous education/training appropriately and effectively. Does not apply rules or strategies blindly. Assesses costs and benefits of alternative solutions and makes timely decisions even with incomplete information.

3. *Exhibits Safety Consciousness.* Follows the details of safety guidelines and instructions. Checks the behavior of others to ensure compliance.

4. *Adapts to Changing Situations.* Is able to maintain commitment when environments, tasks, responsibilities, or personnel change. Does not allow stress in high-pressure situations to interfere with job performance. Easily commits to learning new things when the technology, mission, or situation requires it.

5. *Communicates in Writing.* Communicates thoughts, ideas, and information successfully to others through writing. Uses proper sentence structure including grammar, spelling, capitalization, and punctuation.

6. *Communicates Orally.* Speaks in a clear, organized, and logical manner. Communicates detailed information, instructions, or questions in an efficient and understandable way. Note that this dimension refers to how well the individual can speak and communicate, not whether technical expertise is high or low.

7. *Uses Computers.* Understands and uses computer interfaces and applications (e.g., email, World Wide Web, and Army-specific systems).

8. *Manages Information.* Effectively monitors, interprets, organizes, and redistributes information (i.e., digital, printed, or oral). Does not readily succumb to information overload.

9. *Exhibits Cultural Tolerance.* Demonstrates tolerance and understanding of individuals from other cultural and social backgrounds, both in the context of the diversity of U.S. Army personnel and interactions with foreign nationals during deployments or when training for deployment.

10. *Exhibits Effort and Initiative on the Job.* Demonstrates high effort in completing work. Takes independent action when necessary. Seeks out and willingly accepts responsibility and additional challenging assignments. Persists in carrying out difficult assignments and responsibilities.

11. *Follows Instructions and Rules.* Understands and carries out instructions relayed orally or in writing. Adheres to regulations, policies, and procedures while completing assignments.

12. *Exhibits Integrity and Discipline on the Job.* Maintains high ethical standards. Does not succumb to peer pressure to commit prohibited, harmful, or questionable acts. Demonstrates trustworthiness and exercises effective self-control. Understands and accepts the basic values of the Army and acts accordingly.

13. *Demonstrates Physical Fitness.* Meets Army standards for weight, physical fitness, and strength. Maintains health (e.g., dental hygiene) and fitness to meet requirements, to handle the physical demands of the daily job, and to endure the stress of combat.

14. *Demonstrates Military Presence.* Presents a positive and professional image of self and the Army even when off duty. Maintains proper military appearance. Sets the precedent for other Soldiers to follow.

15. *Relates to and Supports Peers.* Treats peers in a courteous, respectful, and tactful manner. Shows concern for others by providing help and assistance. Backs up and fills in for others when needed.

16. *Exhibits Selfless Service Orientation.* Commits to the greater good of the team or group. Puts organizational welfare ahead of individual goals as required.

17. *Exhibits Self-Management.* Effectively manages own responsibilities (e.g., work assignments, personal finances, family, and personal well being), and appears on duty prepared for work. Sets goals, makes plans, and critically evaluates own performance. Works effectively without direct supervision, but seeks help when appropriate.

18. *Exhibits Self-Directed Learning.* Takes responsibility for mastering skills and learning to apply those skills in the job. As necessary, effectively invests time in learning and practice. Mastery of skills includes those (a) acquired during basic and advanced individual training, and (b) additional skills required by the Soldier's initial assignment.

19. *Demonstrates Teamwork.* Understands own and team tasks in relation to the mission or assignment. Coordinates with and helps members maintain focus on the team's goals.

APPENDIX B

SELECT21 ARMY-WIDE TASKS

Process Casualties

Handle casualties or remains

First Aid

Evaluate a casualty

Perform basic first aid (i.e., CPR, shock prevention, clear throat of casualty)

Administer first aid to wounds to the abdomen or chest

Administer first aid for injuries to extremities or limbs (e.g., put on field dressing, tourniquet, splint)

Administer first aid for an open head wound

Administer first aid for burns or injuries from heat or cold

Administer first aid to CBRN casualty

Transport a casualty

Request medical evacuation

Operate telemedicine transmitting device

Maintenance

Conduct vehicle/FCS platform preventive maintenance checks and services

Mine Installation/Recovery

Locate and neutralize mines

Install antipersonnel mines

Navigate

Navigate using a compass, a map, and overlays

Navigate from one point on the ground to another point

Navigate using electronic or digital tools (e.g., global positioning system receivers)

Prepare field-expedient maps or overlays

Survive

React to combat situations (e.g., attack, ambush, direct/indirect fire) based on training, experience, and own judgment

Communicate by tactical voice or audio systems (e.g., tactical radio, tactical telephone)

Report information of potential intelligence value (SALUTE)

Prepare unit equipment for movement

Select, construct, and camouflage an individual fighting position

React to hazardous incidents (e.g., unexploded ordinance, hazardous materials) based on training, experience, and own judgment

Move through the battlefield, around obstacles, under fire, day or night using visual, hand, or arm signals

Camouflage yourself and your personal equipment

Camouflage equipment (other than personal)

Employ hand-to-hand techniques
Conduct guard duty
Control entry into restricted areas
Conduct a defense by a squad-sized unit
Visually identify vehicles and aircraft (friend and foe)
Establish an observation post
Control or evacuate crowds/non-combatants
Operate a vehicle in a convoy
Defend against air attack
Prevent subversion/espionage directed against the Army

CBRN

Protect yourself and others from NBC injury/contamination using appropriate gear and/or mask
Protect yourself from hazards (e.g., depleted uranium)
Decontaminate yourself or individual equipment using decontamination kits
React to a nuclear, chemical, or biological attack or hazard based on training, experience, and own judgment
Detect or monitor chemical/biological agents using kits, papers, or monitoring devices
Detect radiation and measure dose using detection and measurement tools
Cross a contaminated chemical/nuclear area
Prepare for a friendly nuclear attack

Weapons

Operate personal weapon
Engage targets with personal weapon
Operate squad or crew-served weapon
Engage targets with squad or crew-served weapon
Maintain personal weapon
Conduct safety checks on personal weapon
Maintain squad or crew-served weapon
Operate anti-armor weapon
Engage targets with anti-armor weapon
Conduct safety checks on squad or crew-served weapon
Maintain anti-armor weapon
Conduct safety checks on anti-armor weapon
Locate a target by grid coordinates
Prepare a range card

APPENDIX C

MOS-SPECIFIC TASK CATEGORIES FOR TARGET MOS

Infantryman (11B)

1. Perform General Communications Functions
2. Prepare, Install, and Operate Radios
3. Perform Tactical Operations
4. Perform General Navigation Functions
5. Operate and Maintain Night Vision Devices
6. First Aid
7. Operate and Maintain the Infantry Fighting Vehicle (IFV)
8. Operate and Maintain Weapons (M9, M16 Series, M203, M240 Series, M257, MK19, M249, M60, .50 M2 Machine Gun, M242, M4)
9. Operate and Maintain Antitank Weapons (M136 Launcher, M220, Javelin)
10. Perform General Weapons Functions and Operations
11. Operate Hand Grenades/Mines/Pyrotechnics

Cavalry Scout (19D)

1. Operate in a Net-Centric Environment [includes use of existing communications equipment and the FBCB2 system]
2. Prepare for and React to CBRN Threats
3. Perform Tactical Operations and Functions
4. Perform Mine and Demolition Functions and Operations
5. Operate and Maintain Night Vision Devices
6. Operate and Maintain Weapons (M9, M4, M16A1/M16A2, M203, MK19, M249, M240B, .50 M2 Machine gun)
7. Operate and Maintain M47/M136/M220 Antitank Weapons
8. Operate and Maintain Military UAV/UGV/Robotics
9. Perform HMMWV Functions and Operations
10. Perform Bradley Fighting Vehicle (BFV) Functions and Operations
11. Perform General Skills

Armor Crewman (19K)

1. Operate in a Net-Centric Environment [to include operation of current radio equipment and the FBCB2]
2. Perform Tank Driver Functions and Operations
3. Perform Tank Loader Functions and Operations
4. Operate and Maintain Tank-Mounted Machine Guns
5. Perform Tank Recovery and Towing Operations
6. Perform Tank-Mounted Mine Clearing Equipment Services and Functions
7. Perform Tank Maintenance Functions
8. Maintain, Load, and Stow Tank Gun Ammunition
9. Perform General Tank Crew Operations
10. Perform Tank Gunnery [LOS and NLOS]
11. Operate and Maintain Military UAV/UGV/Robotics

Signal Support Systems Specialist (31U)

1. Maintain Test, Measurement, and Diagnostic Equipment (TMDE)
2. Install, Configure, and Troubleshoot Commercial-Off-the-Shelf (COTS) Equipment
3. Install, Troubleshoot, and Maintain Tactical Computers
4. Install, Troubleshoot, and Maintain Very High Frequency Radios
5. Operate Retransmission Stations (RETRANS) and EPLRS Network Management (ENM) System
6. Install, Troubleshoot, and Maintain Tactical Satellite Equipment
7. Maintain and Troubleshoot Communications Systems for Continuous Operations
8. Restore Communications Security Equipment to Operation
9. Install, Troubleshoot, and Maintain High/Ultra High Frequency Radios
10. Install, Troubleshoot, and Maintain Mobile Subscriber Equipment (MSE)
11. Explain to Operators Proper Use of Equipment
12. Share Critical Information with Peers and Supervisors
13. Identify Potential Threats to System Security

Information Systems Operator/Analyst (74B)

1. Prepare and Maintain Hardware/Software
2. Perform Operations on the Automated Information System (AIS)
3. Process Job Requests
4. Maintain Systems Security
5. Perform Systems Operation Functions

Intelligence Analyst (96B)

1. Perform Map Operations
2. Secure Information and Materials
3. Manage Collection of Intelligence Information
4. Perform Reporting Duties
5. Disseminate Intelligence Information
6. Assist in Intelligence Preparation of the Battlefield
7. Develop Targets
8. Maintain Intelligence Materials

APPENDIX D

PRE-ENLISTMENT KNOWLEDGE, SKILLS, AND ATTRIBUTES

Cognitive Attributes

1. *Oral Communication Skill.* Speaks in a clear, organized, and logical manner. Communicates information or asks questions in an efficient and understandable way. Adapts communication styles to different situations. Uses nonverbal gestures to supplement and reinforce spoken messages.

2. *Oral and Nonverbal Comprehension.* Listens to and comprehends instructions and other related messages. Pays attention to nonverbal cues to help clarify/interpret messages. Asks questions as appropriate.

3. *Written Communication Skill.* Communicates thoughts, ideas, and information successfully to others through writing. Uses proper sentence structure including grammar, spelling, capitalization, and punctuation.

4. *Reading Skill/Comprehension.* Reads and understands written instructions, basic textbooks, and other related written material.

5. *Basic Math Facility.* Knows and applies addition, subtraction, multiplication, division, and simple mathematical formulas. Has the ability to read and interpret various types of graphs and figures (e.g., Cartesian planes).

6. *General Cognitive Aptitude.* The capacity to understand and interpret information that is being presented, the ability to identify problems and reason abstractly, and the capability to learn new things quickly and efficiently.

7. *Spatial Relations Aptitude.* The degree to which an individual can mentally visualize the relative positions of objects in two-dimensional or three-dimensional space, and how they will be positioned if they are moved or rotated in different ways.

8. *Vigilance.* The degree to which an individual can detect infrequent, simple signals over prolonged periods of time without rest.

9. *Working Memory.* The degree to which an individual can maintain information in memory such as words, numbers, pictures, and procedures for short periods of time and to retrieve it accurately.

10. *Pattern Recognition.* The degree to which an individual can detect a known figure or form that is only partially presented or hidden in distracting material.

11. *Selective Attention.* The degree to which an individual can concentrate while performing a task over a period of time without becoming distracted.

12. *Perceptual Speed and Accuracy.* The degree to which an individual can recognize and interpret visual information quickly and accurately, particularly with regard to comparing similarities and differences among words, numbers, objects, or patterns, when presented simultaneously or one after the other.

Temperament Attributes

13. *Team Orientation.* The degree to which an individual identifies with the team and other team members and works to boost team morale and increase the team bond.

14. *Agreeableness.* The degree of pleasantness versus unpleasantness exhibited by an individual in interpersonal relations. Is tactful, helpful, and not defensive, versus touchy, defensive, alienated, and generally contrary.

15. *Cultural Tolerance.* The degree to which an individual demonstrates tolerance and keeps an open mind with respect to individuals from other cultural and social backgrounds.

16. *Social Perceptiveness.* The degree to which an individual is aware of others' reactions and tries to understand why they react the way they do.

17. *Achievement Motivation.* The degree to which an individual sets high standards and strives for accomplishment.

18. *Self-Reliance.* The degree to which an individual depends upon his/her own abilities to overcome difficult or severe situations. Is confident in own abilities. When put in situations that require independent thinking or actions, is able to act appropriately.

19. *Affiliation.* The degree of sociability an individual exhibits. Is outgoing, participative, and friendly versus shy and reserved.

20. *Potency.* The degree of impact, influence, and energy that an individual displays. Is forceful, persuasive, optimistic, and vital versus lethargic and pessimistic.

21. *Dependability.* An individual's characteristic degree of conscientiousness. Is disciplined, well organized, planful, and respectful of laws and regulations, versus unreliable, rebellious, and contemptuous of laws and regulations.

22. *Locus of Control.* An individual's characteristic belief in the amount of control he/she has or people have over rewards and punishments.

23. *Intellectance.* The degree of openness to new experiences and culture an individual possesses and displays. Is imaginative, quick-witted, curious, socially polished, and independent minded versus artistically insensitive, unreflective, and narrow.

24. *Emotional Stability.* The degree to which an individual acts rationally and displays a generally calm, even mood. Typically maintains composure and is not overly distraught by stressful situations.

Physical Attributes

25. *Static Strength.* The ability to exert maximum muscle force to lift, push, pull, or carry objects.

26. *Explosive Strength.* The ability to use short bursts of muscle force to propel oneself (as in jumping or sprinting), or to throw an object.

27. *Dynamic Strength.* The ability to exert muscle force repeatedly or continuously over time. This involves muscular endurance and resistance to muscle fatigue.

28. *Trunk Strength.* The ability to use abdominal and lower back muscles to support part of the body repeatedly or continuously over time without "giving out" or fatiguing.

29. *Stamina.* The ability to maintain physical exertion over long periods of time without getting winded or out of breath.

30. *Extent Flexibility.* The ability to bend, stretch, twist, or reach out with the body, arms, and/or legs.

31. *Dynamic Flexibility.* The ability to quickly and repeatedly bend, stretch, twist, or reach out with the body, arms, and/or legs.

32. *Gross Body Coordination.* The ability to coordinate the movement of the arms, legs, and torso together in activities where the whole body is in motion.

33. *Gross Body Equilibrium.* The ability to keep or regain body balance to stay upright when in an unstable position.

Sensory Attributes

34. *Visual Ability.* The degree to which an individual, with or without corrective lenses, can see details at a distance, discriminate between different colors, see under low light conditions, see objects or movements of objects to his/her side when eyes are focused forward, judge which of several objects is closer or farther, and see objects in the presence of glare or bright lighting.

35. *Auditory Ability.* The degree to which an individual can detect or tell the difference between sounds that vary over a broad range of pitch and loudness, focus on a single source of auditory information in the presence of other distracting sounds, and tell the direction from which sounds originate.

Psychomotor Attributes

36. *Multilimb Coordination.* The ability to coordinate the movements of a number of limbs simultaneously.

37. *Rate Control.* The ability to time continuous anticipatory motor adjustments relative to changes in speed and direction of a continuously moving target or object.

38. *Control Precision.* The ability to make rapid, precise, highly controlled, but not overcontrolled, movements necessary to adjust or position a machine control mechanism (e.g., rudder controls). Control precision involves the use of larger muscle groups, including arm-hand and leg movements.

39. *Manual Dexterity.* The ability to skillfully, engage in well-directed arm-hand movements in manipulating fairly large objects under speeded conditions.

40. *Arm-Hand Steadiness.* The ability to make precise arm-hand positioning movements where strength and speed are minimized; the critical feature is the steadiness with which movements must be made.

41. *Wrist, Finger Speed.* The ability to make rapid discrete movements of the fingers, hands, and wrists, such as in tapping a pencil on paper.

42. *Hand-eye Coordination.* The ability to make precise movements under highly speeded conditions, such as in placing a dot in the middle of a circle, repeatedly, for a page of circles.

Procedural Knowledge and Skill

43. *Basic Computer Skill.* Uses personal computers and software programs. Creates and maintains computer files. Locates and uses information on the Internet and uses other Internet functions including e-mail.

44. *Basic Electronics Knowledge.* Knows general information regarding electronics principles and electronics equipment operation and repair.

45. *Basic Mechanical Knowledge.* Knows general information regarding mechanical principles, tools, and mechanical equipment operation and repair.

46. *Self-Management Skill.* Uses appropriate strategies to self-manage the full range of personal responsibilities (e.g., goal setting, allocation of effort and personal resources, self-assessment of degree of goal accomplishment, and seeking help and advice from others when appropriate).

47. *Self-Directed Learning and Development Skill.* Has a clear goal of maintaining continuous learning. Is proficient at determining learning needs, planning experiences to meet them, and evaluating personal success.

48. *Sound Judgment.* Makes decisions or solves problems in ways that promote outcomes that are effective and rational.

APPENDIX E

ARMY-WIDE FUTURE CONDITIONS

Overview

In the future Army, Soldiers will be working in an environment that will include the following elements:

- A digitized environment in which most training will be provided as needed in the unit rather than at the schoolhouse.
- Soldiers will have much more individual responsibility for keeping pace with changing operational requirements, new technologies, common weapons platforms, and evolving doctrines.
- Future conflicts are expected to involve intense and sustained operations that will require physical and mental stamina to conduct high paced operation over long periods.

Soldiers will be more widely dispersed, working under increased time pressure, and will need to be able to perform tasks with less back-up from supervisors and/or other Soldiers. Following are the descriptions of the four anticipated conditions anticipated to be present in the Future Force.

Condition A: Individual Pace and Intensity

Future conflicts are expected to involve intense and sustained operations that will require physical and mental stamina to conduct high paced operation over long periods. Conditions, such as rules of engagement, hostile forces, threat intent, and force mission, could change daily. Soldiers might go from a peacetime CONUS environment to full combat activities in a matter of a few days. Soldiers must be capable of cycling between periods of work and rest instantaneously and at unpredictable intervals. Soldiers will need to maintain focus and commitment when environments, tasks, responsibilities, or personnel change. Soldiers must recognize and respond to mental cues and images (such as icons and graphics) rather than real life sound or visual images. Soldiers will be required to process information and data flow without becoming overwhelmed, even when tired or stressed. Soldiers will face a greater variety of tasks as a result of missions and operational environments.

Condition B: Learning Environment

Currently, Soldiers are trained at the institution (schoolhouse) using standard materials and methods. In the future, much training will be mission-based and there will be a greater requirement for Soldiers to learn as they go. Operating under considerable time pressure, Soldiers will have to learn material that could be more complex than current materials. Increasingly, Soldiers will be responsible for monitoring their task proficiency and taking steps to acquire needed skills on their own initiative. Soldier training will occur in the unit and will include fellow Soldiers, supervisors, leaders and operational equipment.

Future Combat Systems will include software that allows Soldiers to learn on the actual equipment they will use. Training will take place at or close to operational settings. Soldiers will be required to make frequent use of distance learning and other methods of computer-assisted instruction.

Condition C: Disciplined Initiative

Future Force Soldiers will enter and be assimilated into the Army operational culture much as they are now except at a much more accelerated pace. Soldiers will be spread out both physically and operationally. As a result, they will be required to function with much less direct supervision, interaction, and support from other Soldiers. Soldiers will have to rely less on supervisors and/or other Soldiers to perform assigned tasks. While Soldiers will not be faced with complex decision making beyond their defined responsibilities, they will need to be able to decide what to do with less direct contact and back-up from supervisors and/or other Soldiers. Soldiers with just weeks or months in a unit will be expected to perform, make decisions, initiate and complete actions, and accept responsibilities that currently are performed by specialists and some junior NCOs with 2 to 3 years experience. Soldiers will be required to take initiative in all aspects of their performance including learning new skills, personal and family responsibility, self discipline, and accountability and responsibility in job performance

Condition D: Communication Method and Frequency

Soldiers will be connected electronically with their command chain and with other Soldiers within their unit, at all times and under all conditions. This link will allow voice communication, position location and reporting, data transmission, and visual imagery transmission. Soldiers at all levels, including service support and administrative support, will have the ability to know the common operational picture and be aware of the broader situation. Soldiers must be able to function based on digitized communication (i.e., text, voice, video) instead of face-to-face communication with their supervisors and fellow Soldiers. Soldiers will receive information from multiple sources, that will make additional demands on their attention.

CLOSE COMBAT CLUSTER FUTURE CONDITIONS

Condition A: More Variety in Weapons, Communications, and Vehicle Systems

Technology will increase the capabilities of Future Force weapons, communications, and vehicle systems. Soldiers must be able to use a variety of aiming systems including thermal sights, daylight sights, close combat optics, lasers, and target detection devices. Soldiers will use a range of communication devices including FBCB2, tactical laptops, satellite communications, and dismounted location equipment. As Soldiers move from one assignment to the next, they may need to adapt to different combat vehicle and fighting systems. For example, the same Soldiers might crew M1-tank variants, Strykers, and FCS troop transport/fighting vehicles. Soldiers will be required to interact with and operate unmanned aerial vehicles (UAVs) and unmanned ground vehicles (UGVs).

Condition B: Deployment in Different Configurations

Most of the time Soldiers will be deployed with members of their own unit, but sometimes they may be part of teams made up of individuals from other branches, militaries, and/or agencies. The personnel assigned to these teams will depend greatly on the specific missions. Deployment to almost any part of the world will require Soldiers to interact with military personnel from other nations and to have contact with indigenous personnel. Soldiers must be able to communicate with and interact with members from all branches of the US military, soldiers from other countries, and civilian specialists.

Condition C: Changes in Tasks

Soldiers will fill different roles, depending on the mission, theater of operation, or team assignment. The tasks the unit will be required to perform are likely to change frequently and special conditions could continue for months at a time. Because the level of operational intensity will vary, Soldiers must be able to stand down from the rush of combat and assume a lower level of intensity. They also may need to rapidly escalate into a combat mode when required. Soldiers will interact with indigenous personnel in a variety of situations (e.g., giving aid and assistance, patrolling neighborhoods, working at checkpoints). Soldiers will be required to maintain a professional attitude during these interactions, while remaining aware of their surroundings and alert to potential danger. Soldiers may be fighting with a population one week and providing support services to these same people the next week. As tasks and rules of engagement change, it will be important that Soldiers be able to adjust their attitudes toward those once regarded as enemies.

31U/74B CLUSTER FUTURE CONDITIONS

Condition A: Changing Breadth and Depth of Knowledge

The percentage of vehicles and individuals that have communication devices will increase. Soldiers in communications support jobs will be responsible for more pieces and a larger variety of equipment, but there will also be less need for technical depth in any particular area. Equipment will be modularized, so Soldiers will replace modules or send the item away to be repaired. Some of the expectations include:

Condition B: Broader Areas of Responsibility

Communications support Soldiers will be cross-trained in a variety of signal technologies. Commercial-off-the-shelf (COTS) equipment and other common software and hardware will become more widely used. Soldiers will be given a generic grounding in communications and information technology and management, but all operational familiarity on specific systems, software, and capability will be acquired at the unit of assignment.

Condition C: Increased Operator Support

Because of a wide variety of technologies and a larger number of users/operators, Soldiers in communications support jobs will need to interact frequently with users/operators. Communications support Soldiers may or may not be physically located with the unit they are tasked to support, so they will perform many of their functions remotely. Some of the expectations include:

APPENDIX F

FAKING RESEARCH INSTRUCTIONS TO RESPONDENTS

The following instructions to respondents apply to these instruments:

Work Values Inventory (WVI) (original version)
Work Suitability Inventory (WSI)
Work Preferences Survey (WPS)
Interest Finder Questionnaire (IFQ)

Fake Maximum Condition

We want you to answer the items on this measure in a way that you think will make you look as good to the Army as possible. In fact, it is OK if your responses do not accurately describe how you really are—we want to know how good you can make yourself look. To help you get the right frame of mind, pretend you are an applicant who is willing to say anything to get into the Army.

Fake with Lie Scale Condition (WSI, WPS, IFQ)

We want you to answer the items on this measure in a way that you think will make you look as good to the Army as possible. In fact, it is OK if your responses do not accurately describe how you really are—we want to know how good you can make yourself look. To help you get the right frame of mind, pretend you are an applicant who is willing to say anything to get into the Army.

You need to be aware, however, that this measure contains a "lie" scale designed to identify people who make themselves look better than they are. Thus, although we want you to respond to the items in a way that makes you look as good as possible to the Army, try to do so in a way that doesn't make it obvious you are lying.

Coaching Condition (WVI only)

We want you to answer the items on this measure in a way that we think will make you look as good to the Army as possible. Here's exactly how we would like you to do it. For each item, indicate the statement that sounds "most like the Army" is MOST important to you on your ideal job, and indicate the statement that sounds "least like the Army" is LEAST important to you on your ideal job. It is OK if your responses do not indicate your true preferences. We want to see if you can find the statements that are most and least like the Army.

The Rational Biodata Inventory (RBI) was administered under two conditions using instructions intended to simulate faking that might actually occur under operational conditions. One is coached and the other is not.

<u>RBI Fake Operational</u>

When taking this test, imagine that the results are very important to you because they affect your chances of joining the Army and getting the MOS that you want. The best thing to do is to pretend that you've just finished taking the ASVAB and now are taking a second test that is as important for getting into the Army as the ASVAB is.

<u>RBI Fake Operational with Coaching</u>

When taking this test, imagine that the results are very important to you because they affect your chances of joining the Army and getting the MOS that you want. The best thing to do is to pretend that you've just finished taking the ASVAB and now are taking a second test that is as important for getting into the Army as the ASVAB is.

Here are some tips for scoring well on the test. For each item, try to respond in a way that will make you look as good as possible to the Army. This means circling item options that seem to say good things about you and not circling items that seem to say bad things about you. However, be careful not to exaggerate so much that it becomes clear to the Army that you are not telling the truth.

The Predictor Situational Judgment Test (PSJT) was administered under two conditions. Note that these are not actually "faking" conditions since the PSJT is a knowledge-based measure rather than a self-report temperament measure. We did, however, want to examine the impact of coaching that is a potential threat to the accuracy of scores on this instrument in an operational setting.

PSJT Uncoached Condition

Pretend that you are applying to join the Army. You very much want to be accepted into the Army. Your score on this test will be used to determine whether you are accepted. Answer the questions the way you would under these conditions.

PSJT Coached Condition

This time when you take the test, it is important that you try to get the highest score possible. Here are a few tips for getting a high score. In general, you should give **higher** ratings to actions that:

- Get things done
- Consider the needs and feelings of other people
- Maintain/improve morale, or
- Show integrity and honesty

In contrast, you should give **lower** ratings to actions that:

- Avoid exerting effort
- Ignore other persons' needs and feelings
- Hurt morale
- Involve deception, or show a lack of integrity

These coaching tips are shown in your handout. Use them to help you make your ratings.

APPENDIX G

JOB KNOWLEDGE TEST BLUEPRINTS

Army-Wide Test

11B (Infantryman) Test

31U (Signal Support Systems Specialist) Test

Army-Wide Test Blueprint

Code		% of test
1	*First Aid*	20
	1a Evaluate a casualty	
	1b Perform basic first aid (i.e., CPR, shock prevention, clear throat of casualty)	
	1c Administer first aid to wounds to the abdomen or chest	
	1d Administer first aid for injuries to extremities or limbs (e.g., put on field dressing, tourniquet, splint)	
2	*Vehicle Maintenance*	6
	2a Conduct vehicle/Future Combat System (FCS) platform preventive maintenance checks and services	
3	*Mine Installation and Mine Recovery*	5
	3a Install antipersonnel mines	
	3b Locate and neutralize mines	
4	*Navigate*	15
	4a Navigate from one point on the ground to another point	
	4b Navigate using a compass, a map, and overlays	
	4c Prepare field-expedient maps or overlays	
5	*Survive*	20
	5a Select, construct, and camouflage an individual fighting position	
	5b Establish an observation post	
	5c React to hazardous incidents (e.g., unexploded ordinance, hazardous materials) based on training, experience, and own judgment (not CBRN)	
	5d Report information of potential intelligence value (SALUTE)	
	5e Communicate by tactical voice or audio systems (e.g., tactical radio, tactical telephone)	
6	*CBRN (Chemical, Biological, Radiological, & Nuclear)*	14
	6a Protect yourself and others from NBC injury/contamination using appropriate gear and/or mask	
	6b Protect yourself from hazards (e.g., depleted uranium)	
	6c Decontaminate yourself or individual equipment using decontamination kits	
7	*Weapons (M16, M203, M60, M249)*	20
	7a Engage targets with personal weapon	
	7b Engage targets with squad or crew-served weapon	
	7c Operate personal weapon	
	7d Operate squad or crew-served weapon	
	7e Conduct safety checks on personal weapon	
	7f Maintain personal weapon	
	7g Maintain squad or crew-served weapon	
Total		**100**

11B Test Blueprint		
Code		% of test
1	***Perform General Communications Functions***	**5**
	1a Use Visual Signaling Techniques	
2	***Prepare, Install, and Operate Radios***	**5**
	2a Operate in Radio Nets (similar to Perform Voice Communications)	
3	***Perform Tactical Operations***	**25**
	3a Move as a Member of a Fire Team	
	3b Prepare Positions for Individual and Crew-Served Weapons During an Urban Operation	
	3c Select Hasty Firing Positions During an Urban Operation	
	3d Estimate Range	
4	***Perform General Navigation Functions***	**15**
	4a Read and Navigate with a Map and a Protractor	
	4b Navigate with a Compass; Determine a Grid Azimuth Using Compass	
5	***Operate and Maintain Night Vision Devices***	**5**
	5b Maintain and Operate Night Vision Goggles	
6	***Operate and Maintain Weapons (M16 Series, M203, M240 Series, M249)***	**25**
	6a[1] Operate M240	
	6b Correct Malfunctions of a Weapon	
	6c Zero a Weapon	
	6d Zero an Aiming Light to a Weapon	
7	***Operate and Maintain Antitank Weapons (M136 Launcher)***	**5**
	7a[2] Operate and maintain M136 Launcher	
8	***Perform General Weapons Functions and Operations***	**5**
	8a Prepare an Anti-armor Range Card	
9	***Operate Hand Grenades***	**5**
	9a Employ Hand Grenades	
Note:	The following category will only be given to Soldiers in units that use Bradley Fighting Vehicles	
10	***Operate and Maintain the Infantry Fighting Vehicle (IFV)***	**5**
	10a Maintain the Hull on an IFV	
	10c Drive an IFV	
Total		**100**

[1] 6a is a combination of three tasks related to operating an M240 machine gun, including engaging targets, maintaining weapon, and performing a function check

[2] 7a is a combination of two tasks related to operating an M136 Launcher, including engaging targets and performing misfire procedures

31U Test Blueprint

Content Area		% of test
1	***Maintain Test, Measurement, and Diagnostic Equipment (TMDE)***	6
2	***Install, Configure, and Troubleshoot Commercial-Off-the-Shelf (COTS) Equipment***	6
	2a Install Network Hardware/Software in a Desktop/Laptop IBM or Compatible Microcomputer (e.g., Windows, Unix, FBCB2, Solaris)	
	2b Troubleshoot a Desktop/Laptop IBM or Compatible Microcomputer	
3	***Install, Troubleshoot, and Maintain Tactical Computers***	16
	3a Install a Tactical Local Area Network (LAN)	
	3b Troubleshoot a Tactical Local Area Network (LAN)	
	3c Perform Scheduled Unit Level Maintenance (ULM) on Common Hardware and Software (CHS) Within a Standardized Integrated Command Post System (SICPS)	
	3d Install Force XXI Battle Command Brigade and Below (FBCB2)	
4	***Install, Troubleshoot, and Maintain Very High Frequency Radios***	11
	4a Troubleshoot Secure ASIP, SIP and PRC-140 Radio Sets with or without the AN/VIC-1 or AN/VIS-3	
	4b Troubleshoot Single Channel Ground And Airborne Radio Systems (SINCGARs) ICOM with or without AN/VIC-1 or AN/VIS-3	
	4c Install Single-Channel Ground And Airborne Radio Systems (SINCGARs) ICOM with or without the AN/VIC-1 or AN/VIS-3	
	4d Install Secure ASIP, SIP and PRC-140 Radio Sets with or without the AN/VIC-1 or AN/VIS-3	
5	***Operate Retransmission Stations (RETRANS)***	8
	5a Operate Secure AN/VRC-92 RETRANS	
	5b Operate Secure Tactical Satellite (TACSAT) RETRANS Using AN/PSC-5	
6	***Install, Troubleshoot, and Maintain Tactical Satellite Equipment***	6
	6a Troubleshoot Secure Tactical Satellite Communications (TACSAT) Radio Set AN/VSC-117 or a Similar TACSAT Radio Set	
	6b Install Secure AN/VSC-117 or a Similar TACSAT Radio Set	
7	***Maintain and Troubleshoot Communications Systems for Continuous Operations***	21
	7b Install, Maintain, and Operate Generators	
	7c Install, Maintain, and Operate Power Supply	
	7d Perform Scheduled Unit Level Maintenance (ULM) On Antenna Group OE-254/GRC or a Similar Antenna System	
	7e Perform Unit Level Maintenance (ULM) On Communications Equipment Within Standardized Integrated Command Post System (SICPS)	
8	***Restore Communications Security Equipment to Operation***	11
9	***Install, Troubleshoot, and Maintain High/Ultra High Frequency Radios***	6
	9a Install Improved High Frequency Radio (IHFR) Set AN/GRC-213 or a Similar System	
	9c Troubleshoot Enhanced Position Location Reporting System (EPLRS) Radio Set AN/VSQ-2(V)1/(V)2	
	9d Install Enhanced Position Location Reporting System (EPLRS) Radio Set AN/VSQ-2(V)1/(V)2	
10	***Install, Troubleshoot, and Maintain Mobile Subscriber Equipment (MSE)***	9
	10a Install Mobile Subscriber Radiotelephone Terminal (MSRT) AN/VRC-97	
	10b Troubleshoot Mobile Subscriber Radiotelephone Terminal (MSRT) AN/VRC-97 System	
Total		**100**

APPENDIX H

DETAILED RESPONSES TO REPETE WRITE-IN QUESTIONS

Table H1. Content Analysis of Write-In Responses for Computer Courses

n	Course	Course Content
227	Computer Applications I / Computer Fundamentals	Introductory information about computer hardware and software, disk maintenance and working with folders and files. Introductory training in MS Windows, Word, Excel, and PowerPoint and internet usage.
86	Keyboarding, Typing, and Data Entry I	Mastery of alphabetic and numeric keyboard using touch system. Formatting, speed, and skill development.
42	Word Processing	Provides in-depth knowledge and skill using word processing software such as Word
35	Management/Business Computer Information Systems	Introductory course for business or computer science majors. Covers the fundamentals of computer information systems and gaining an understanding of fundamental programming concepts. Introduces algorithm design, logic diagrams, coding, and debugging.
25	Computer Science I	Introduction to computer information systems for computer science students. Covers the basic architecture and design of digital computers and the software that runs on them.
24	Computer Applications II	Training in software for office applications such as Word, Excel, Access, and PowerPoint and integration of applications.
21	Microcomputer Hardware/Software Maintenance, Repair, and Support	In-depth coverage of PC hardware and operating systems (e.g., DOS, Windows) and the applications they run. Covers knowledge required by CompTIA for A+ certification.
21	Web Page Design with HTML/JavaScript	Students learn to crate attractive and appealing Web pages and get them quickly running on the Web.
18	Computer Programming Basics	Introduction to the fundamental principles of programming and to different programming paradigms. Students develop reasoning skills needed for programming.
18	Network Administrator/Associate Course 1	Covers electrical safety, network terminology and protocols, network standards, local area networks, wide area networks, open system interconnection models, physical cabling, cabling tools, routers, and TCP/IP.
15	Computer Graphics: Basics	Uses industry standard software (e.g., Adobe Illustrator) in the application of basic design principles.
13	Computer-Aided Design/Drafting	Uses AutoCAD software. Covers drawing setup, drawing, editing, drawing text, and dimension practices.
10	Desktop Publishing I	Introduction to desktop publishing software. Includes basic design, layout, selection of type and illustration for in-house publishing.
10	Spreadsheets: Microsoft Excel	Covers electronic spreadsheet concepts, software, and problem solving strategies.
7	Computer Graphics: Digital Design and Imaging	Teaches skills and techniques in the creation and manipulation of images.
7	Computer Programming: C/C++	Programming using C and C++ languages, emphasizing program development and design, debugging techniques and common basics of the C/C++ languages.

Table H1. (Continued)

n	Course	Course Content
6	Computer Science II	Covers advanced computer science topics regarding machine architecture and algorithm development.
6	Microsoft PowerPoint	Topics include creating a presentation using a design template and auto layouts; creating a presentation on the Web; using embedded visuals; and creating a self-running presentation using animation.
5	Visual Communication	Covers tools available to designer. Use of industry software applications for real world projects that provide layout experience.
4	Database Use and Design	Introduction to database application software. Emphasizes the use of the computer as a tool in a variety of personal and business environment.
4	Internet Fundamentals	Hands-on instruction in the use of the Internet and World Wide Web. Covers software tools and techniques used to search, retrieve and create internet documents.
4	Microsoft Windows	Installation, configuration, and deployment of the Windows operating system.
4	Visual Communication: Media Design	Covers interaction of type, image, motion, sound, and sequence in staging for various media formats including commercials.
3	Computer Programming: Java	Introduction to object-oriented programming using Java. Covers how to design, code, and debug Java applications and applets.
2	Computer Programming: Numerical Analysis	Introduction to numerical algorithms for computer scientists. Covers topics such as the effect of finite precision arithmetic and basic numerical methods.
2	Computer Programming: Visual Basic	Introduction to object-oriented programming using Visual Basic. Emphasis is on program development and design, application of logic, debugging, basics of syntax.
2	Visual Communication: Digital Video	Introduction to digital video, digital audio, presentation graphics, and multimedia applications emphasizing technical and aesthetic fundamentals of sequential imaging.
1	Accounting Computer Applications	Use of popular commercial accounting software. Use of spreadsheets and other appropriate software for report preparation.
1	Computer Graphics: 2D Animation and Video	Studies 2D computer-based animation mixed with video for expressive images.
1	Computer Graphics: 3D Modeling and Rendering	Studies 3D computer-based modeling and rendering. Covers planar views using polygon construction of objects with numeric input.
1	Computer Programming: Assembly Language	Computer programming using an assembly level language. Includes computer internal structure and addressing, data representation codes, number systems, machine instruction formats, etc.
1	Computer Programming: COBOL	Computer programming using COBOL language within the mainframe environment. Batch programs developed based on given program specifications involving sequential/entry sequence files.
1	Computer Programming: Fortran	Computer programming using the Fortran language.
1	Desktop Publishing II	Covers advanced applications of desktop publishing software.
1	Keyboarding, Typing, and Data Entry II	Advanced formatting, speed, and skill development in keyboarding.

Table H2. Content Analysis of Write-In Responses for Computer Certifications

n	Response Content	Example Certifying Organization
8	Computer Fundamentals	Brainbench
7	A+	Computing Technology Industry Association (CompTIA)
5	MS Excel Fundamentals	Brainbench
5	MS Word Fundamentals	Brainbench
3	Auto CAD	Society of Automotive Engineers
3	Cisco Certified Network Professional	Cisco
2	Microsoft Certified Professional	Microsoft
2	Microsoft Certified Systems Engineer	Microsoft
2	Network + Certified Service Technician	Computing Technology Industry Association (CompTIA)
2	Networking Concepts	Brainbench
2	MS Office Specialist	Microsoft
2	Web Design Concepts	Brainbench
1	Electronic Switching System (ESS)	Enterasys
1	IBM Authorized Repair Center Technician	IBM
1	Information Technology Terminology	Brainbench
1	Microsoft Certified Applications Developer	Microsoft
1	Microsoft Certified Systems Administrator	Microsoft
1	MS Access Fundamentals	Brainbench
1	MS Windows Fundamentals	Brainbench
1	Web Hosting	Brinkster

Table H3. Content Analysis of Write-In Responses for Athletic Certifications

n	Response Content
80	Athletic Achievement Total (see below)
19	Most Valuable Player Award, Letter, or Other Award (no reference to sport)
1	Archery
8	Baseball or Softball
12	Basketball
1	Boxing
1	Canoeing
14	Football
4	Martial Arts
6	Martial Arts 1st or 2nd Degree Black Belt
1	Martial Arts Brown Belt
1	Martial Arts Yellow Belt
9	Physical Fitness
4	Rodeo
8	Soccer
1	Tennis
15	Track and Field
1	Volleyball
3	Wrestling
14	Certified Lifeguard (Red Cross)
6	Certified Personal Trainer or Instructor's Certificate
3	Scuba Diving License or Certification NAUI or PADI
1	Skydiving License

Table H4. Content Analysis of Write-In Responses for Driving and Piloting Certifications

n	Response Content
36	Driver's License: Motor Vehicle Operator
21	Industrial Equipment License or Certification: Forklift
11	Driver's License: Commercial (CDL)
7	Boater's License or Watercraft Operator Certification
5	Pilot's License: Private
2	Other Vehicle Certifications/Licenses: Snowmobile
1	Driver's License: Motorcycle
1	Industrial Equipment License or Certification: Electric Ladder
1	Industrial Equipment License or Certification: Front End Loader
1	Industrial Equipment License or Certification: Heavy Machinery
1	Industrial Equipment License or Certification: PIT
1	Other Vehicle Certifications/Licenses: ATV
1	Pilot's License: Certified Flight Instructor
1	Pilot's License: Commercial

Table H5. Content Analysis of Write-In Responses for Protective Service-Related Certifications

n	Content
6	Firearms: Certification or License
5	Firearms: Hunter's Safety Certification
4	Firearms: Concealed Firearms/Handgun/Weapon Permit
4	Law Enforcement: Student
2	Firearms: Marksmanship Qualification
2	Firefighter program completion
1	Law Enforcement: Certified Correctional Officer

Table H6. Content Analysis of Write-In Responses for Mechanic-Related Certifications

n	Content
4	Certified Autobody Repair Technician
3	Certified Mechanic (ASE)
3	Certified Welder
1	Certified Brakes Technician (ASE)
1	Certified Automotive Systems Technician
1	Certified Medium/Heavy Truck Mechanic

Table H7. Content Analysis of Write-In Responses for Health Care Certifications

n	Content
78	CPR and First Aid
12	Certified Nursing Assistant (CNA)
5	Emergency Medical Technician (EMT)
3	Certified Phlebotomist
1	Cardiac-Related: Defibrulator Certification
1	Cardiac-Related: Life Support
1	Cardiac-Related: Vascular Technician

APPENDIX I

SCORING DETAILS FOR PERSON-ENVIRONMENT FIT MEASURES

Scoring Algorithm for the Work Values Inventory (WVI)

As noted in Chapter 13, for the remainder of Select21, we plan to adopt an algorithm for scoring the WVI scales that parallels the algorithm used to score the Minnesota Importance Questionnaire (MIQ; Gay, Weiss, Hendel, Dawis, & Lofquist, 1971) and the Occupational Information Network (O*NET) Work Importance Profiler (WIP; McCloy, Waugh, Medsker, Wall, Rivkin, & Lewis, 1999). We subsequently refer to this as the MIQ/WIP algorithm. The steps taken to score the WVI using this algorithm are as follows.

1. Convert the rank that respondents assign to each of the 28 reinforcers to "adjusted votes" by subtracting the rank from 28, and then adding 0.5.

Votes represent the number of times a reinforcer is ranked over the other reinforcers in the list. The addition of 0.5 is an adjustment to avoid arriving at a proportion of 0 in Step 7. If this adjustment were not made, then no score would result for the reinforcer that was ranked last (proportions of 0.0 do not have *z*-score equivalents; see Steps 5 and 6).

2. Add 1 to the adjusted number of votes for each reinforcer considered "important" by the respondent.

On the final task of the WVI, respondents distinguish between the reinforcers they feel are important to have on their ideal job, and those they feel are unimportant to have.

3. Establish a "zero-point reinforcer" for each respondent (an imaginary "29th" reinforcer) that falls between the lowest ranked reinforcer that the respondent considers important to have on his/her ideal job, and the highest ranked reinforcer that the respondent considers not important to have on his/her ideal job.

4. Calculate the adjusted number of votes for the zero-point reinforcer by summing the number of reinforcers considered "not important" by the respondent, and adding 0.5.

5. Divide the respondent's adjusted number of votes for each reinforcer by 29 (including the zero-point reinforcer).

6. Convert the proportions calculated in Step 5 into *z*-scores.

The *z*-score for each reinforcer is the point on the standard normal distribution that corresponds to the cumulative density equal to the proportion calculated in Step 5. In other words, the proportion in Step 6 reflects the proportion of the standard normal distribution that falls at or below a given *z*-score.

7. Subtract the z-score of the respondents' zero-point reinforcer from his/her z-scores on all other reinforcers.

This step centers each respondent's set of scores around zero. Reinforcers with positive scores are considered important by the respondent, and reinforcers with negative scores are considered unimportant by the respondent. The resulting scores are the WVI scale scores based on the MIQ/WIP algorithm.

Rescaling the Army Description Inventory (ADI) for Fit Analyses

As noted in Chapter 13, one of the issues we need to confront before comparing subject matter experts' (SMEs) ADI ratings to respondents' WVI ratings is placing them on the same metric. Recall that ADI ratings were made on a 5-point-scale, and WVI scale scores are expressed in a z-score metric based on the MIQ/WIP algorithm. To use the ADI and WVI data to calculate fit indexes and predict Select21 criteria, we rescaled the ADI scores to the same metric as the WVI scores. The following steps are an adaptation of the MIQ/WIP algorithm that transforms the ADI ratings into a z-score metric with a meaningful zero-point.

1. Calculate mean ADI scores for the 28 reinforcers that correspond to the work values assessed on the field test version of the WVI.

2. Convert the mean ADI scores to ranks (e.g., the reinforcer with the highest mean rating is assigned a rank of 1, and the reinforcer with the lowest mean rating is assigned a rank of 28).

3. Convert ranks to "adjusted votes" by subtracting each reinforcer's rank from 28, and then adding 0.5.

The addition of 0.5 (as discussed above) is an adjustment that is used to avoid arriving at a proportion of 0 in Step 7.

4. Establish a "zero-point reinforcer" (an imaginary "29th" reinforcer) that has a level of supply falling between the lowest ranked reinforcer that is considered "generally supplied" by the Army, and the highest ranked reinforcer that is considered "not generally supplied" by the Army.[3]

Given that SMEs did not make an absolute distinction between which reinforcers are "generally supplied" by the Army and which ones are "not generally supplied," establishing a zero-point required a judgment based on the ADI ratings provided. A reinforcer was considered to be generally supplied if (a) it had a mean ADI rating ≥ 3.75, and (b) 65% or more of SMEs gave it a rating of 4 or above.[4] Fourteen reinforcers were considered "generally supplied" for the current Army, with the zero-point reinforcer falling between Societal Contribution (ranked 14th)

[3] We use "generally" because it would be difficult to say that several reinforcers (particularly those in the mid supply category) are either clearly supplied or not supplied.

[4] We made an exception to this rule for the future Army ratings because there was a notable drop in mean ratings between the 16th and 17th ranked reinforcers.

and Leisure Time (ranked 15th), and 16 reinforcers were considered "generally supplied" for the future Army, with the zero-point falling between Supportive Supervision (ranked 16th) and Activity (ranked 17th).

5. Add 1 to the adjusted number of votes for each reinforcer considered "generally supplied."

6. Calculate the adjusted number of votes for the zero-point reinforcer by adding up the number of reinforcers considered "not generally supplied" and adding 0.5.

7. Divide the adjusted number of votes for each reinforcer by 29 (including the zero-point reinforcer).

8. Convert the proportions calculated in Step 7 into *z*-scores.

As discussed earlier, the *z*-score for each reinforcer is the point on the standard normal distribution that corresponds to the cumulative density equal to the proportion calculated in Step 7.

9. Subtract the *z*-score of the zero-point reinforcer from the *z*-scores of all other reinforcers.

This step centers the reinforcers' scores around zero. Reinforcers with positive scores are generally supplied by the Army, and reinforcers with negative scores are not generally supplied by the Army. The resulting scores are the ADI scale scores that we use to compare the similarity of WVI and ADI profiles in Chapter 13, as well as the scores we will use in the future to model the joint effects of WVI and ADI scores on Select21 criteria. Final ADI scale scores are presented in Table I.1.

Combining Person-Side and Environment-Side Scores for Prediction of Criteria[5]

An important distinction we want to make is between describing the level of similarity between person and environment profiles, and how we will combine person and environment data to predict relevant criteria. In Chapter 13, we described the similarity between respondents' profiles on the person-side measures (e.g., WVI, Pre-Service Expectations Survey [PSES]) and Army profiles based on environment-side measures (e.g., ADI, Army Environment Survey [AES]). Specifically, we provided descriptive statistics on two commonly used fit indexes (D^2 and r). These fit indexes provided information on overall profile similarity (which covers elevation, scatter, and shape differences), as well as similarity strictly in terms of shape.

In subsequent validation efforts, our focus will shift to how we can best combine person and environment data to predict various criteria (e.g., satisfaction, attrition). Our goal in combining these data is to maximize prediction of the desired criterion, yet do so in a way that is consistent with P-E fit theory. Although the fit indexes reported in Chapter 13 are useful for describing similarity of profiles, past research has indicated that they put unrealistic and methodologically problematic constraints on person-environment-criterion (P-E-C) relations (Cable & Edwards, 2004; Edwards, 1991, 1993). Past research has also illustrated how by using

[5] Further details on the methods discussed in this section are presented in Putka (2005).

such fit indexes, researchers often fail to fully realize the predictive validity of their person and environment data (e.g., Edwards, 1993; Edwards & Parry, 1993).

Table I1. Final ADI Scale Scores for Use with WVI Scale Scores in Fit Analyses

	Current Army			Future Army		
Dimension	Rank	Mean Rating	Final Scale Score	Rank	Mean Rating	Final Scale Score
Co-Workers	1	4.33	2.11	1	4.67	2.29
Advancement	2	4.30	1.63	2	4.50	1.80
Feedback	3	4.09	1.36	9	4.33	0.72
Emotional Development	4	4.06	1.17	3	4.50	1.54
Achievement	5	4.05	1.01	15	4.00	0.17
Social Service	6	4.00	0.88	10	4.33	0.62
Physical Development	7	3.96	0.76	5	4.50	1.19
Team Orientation	8	3.93	0.65	4	4.50	1.35
Personal Development	9	3.92	0.54	6	4.50	1.05
Fixed Role	10	3.84	0.45	18	3.50	-0.18
Travel	11	3.84	0.35	12	4.17	0.44
Recognition	12	3.78	0.26	11	4.33	0.53
Social Status	13	3.72	0.17	8	4.50	0.82
Societal Contribution	14	3.72	0.09	7	4.50	0.93
Leisure Time	15	3.65	-0.09	19	3.00	-0.27
Leadership Opportunities	16	3.52	-0.17	20	3.00	-0.37
Supportive Supervision	17	3.51	-0.26	16	4.00	0.09
Ability Utilization	18	3.50	-0.35	13	4.17	0.35
Activity	19	3.20	-0.45	17	3.83	-0.09
Esteem	20	3.20	-0.54	14	4.17	0.26
Creativity	21	3.06	-0.65	21	3.00	-0.47
Variety	22	2.98	-0.76	22	3.00	-0.58
Influence	23	2.72	-0.88	24	2.50	-0.84
Comfort	24	2.72	-1.01	25	2.50	-1.00
Flexible Schedule	25	2.65	-1.17	26	2.17	-1.19
Autonomy	26	2.41	-1.36	27	2.17	-1.45
Independence	27	2.38	-1.63	23	2.67	-0.71
Home	28	2.13	-2.11	28	2.17	-1.94

Note. Dimensions are shown in descending order of their mean current Army rating. The correlation between final scale score profiles is .90. ICC(A,1), an index of absolute agreement between the profiles, is equal to .89.

In light of such problems, many researchers have suggested viewing the constraints imposed by fit indexes on P-E-C relations as hypotheses to be tested, and assessing their tenability using more flexible models (Cronbach, 1958; Edwards, 1993; Hesketh & Gardner, 1993; Tinsley, 2000). The most common response has been to use polynomial regression (Edwards, 1991, 1993). Polynomial regression has two distinct advantages over simply using fit indexes as predictors. First, it is advantageous from a theoretical perspective because it allows researchers to assess the viability of the constraints imposed on P-E-C relations by fit indexes. Second, from a practical perspective, it allows researchers to free the aforementioned constraints, and in turn more fully realize the predictive validity of their person and environment data.

Although polynomial regression has benefits over the use of fit indexes to predict criteria, it has limited utility for Select21. Specifically, the approach is most applicable in situations where there is variation in environment-side data across persons. Such variation is not present when one is targeting a person's fit to a single environment (e.g., a single job or organization) and when the environment-side data reflect aggregate SME ratings. As described in Chapter 13, this is exactly the situation we face in Select21. Putka (2005) illustrates how use of polynomial regression in such a situation is problematic because it results in exclusion of environment data from the modeling process.

Modeling Respondents' Fit to the Army Environment

Because of the limitations of fit indexes and polynomial regression for dealing with the situation faced in Select21, we plan to adopt the following two-stage approach for modeling P-E-C relations. The first stage involves modeling the relation between each dimension of fit (e.g., variety, autonomy) and the criterion of interest (e.g., attrition cognitions, attrition). Once the models for each fit dimension have been identified, the second stage involves using terms from each fit dimension's final model to form a single overall model for predicting the criterion. In the sections that follow, we further detail the steps involved in this approach.

1. Assess the fit of a simple linear model without environment-side data

We will begin by fitting a simple linear model for each person-side variable (e.g., WVI Autonomy, WPS Realistic). In these models, the given person-side variable will be the only predictor in the model. The fit of these models will serve as a baseline against which subsequent, more complex models that include environment-side data will be compared. Specifically, this "P-only" model will give us a point of comparison for determining whether environment-side data significantly increment the validity of the model. This is something that is rarely done in the P-E fit literature and has been a source of criticism, as sometimes either person- or environment-side data might not offer any incremental validity beyond the other (Tinsley, 2000). Although seemingly contrary to the P-E fit paradigm, such a model will likely hold if the Army environment provides either a very high or very low level of the attribute in question. In such cases, nearly all the person-side data would fall below/above the level of supply in the environment, and as such, create a situation where misfit occurs in only one direction. If so, a linear model indexing the relation between the person-side variable and the criterion will likely suffice.

2. Assess the fit of models that use common fit indices as predictors

In this step, we will introduce environment-side data into the modeling process, but do so in a way that minimizes the complexity of the model. Specifically, we will fit two models – one that uses $|d|$ as the sole predictor of the criterion, and another that uses d^2 as the sole predictor.

3. Assess the fit of single knot spline models

In this step, we will fit models to the data that relax many of the constraints imposed by the fit indices used in Step 2, yet allow us to control where the change in the relation between the person-side variable (P) and criterion (Y) occurs. This will be done by fitting *single knot spline*

models (linear and quadratic splines) so that the point at which the *P-Y* relationship is hypothesized to change occurs at the point where *P* begins to exceed *E*—the Army's level of supply of a given attribute (e.g., autonomy).

Spline models are a type of non-linear regression model that allow researchers to model differences in the magnitude, direction, and functional form (e.g., linear, quadratic) of predictor-criterion relationship for different ranges of the predictor variable. Splines are particularly useful in the context of P-E fit because they allow one to specify a priori where and how changes in predictor-criterion relations are hypothesized to occur (e.g., respondents' need for Autonomy is expected to become more positively related to attrition once respondents' need for autonomy starts exceeding the Army's supply of autonomy).[6] In Select21, we will use splines to model changes in the relationship between a person's needs/expectations (regarding a given attribute) and a criterion once the person's needs/expectations begin to exceed the Army's supply (*E*) of the attribute. Such models will allow us to relax both the symmetry (i.e., the *P-Y* relation will be equal in magnitude but opposite in sign on either side of *E*) and functional form constraints (i.e., the functional form of the *P-Y* relation will be identical on both sides of *E*—linear in the case of $|d|$, and quadratic in the case of d^2) imposed by $|d|$ and d^2.

In assessing the fit of the single knot spline models to the data and judging their fit relative to models from the previous steps, we will consider not only whether they account for a greater amount of variance in the criterion, but also whether the direction of the regression coefficients that result from the model make sense (e.g., is the direction of the coefficients meaningful given theory surrounding that dimension of fit?). If the direction of the regression coefficients does not conform to expectations, the results may be an artifact of the particular sample, and thus may have limited generalizability. As such, when comparing spline models to the simpler models from Steps 1 and 2, we will consider both the fit of the model and the meaningfulness of the *P-Y* relations that the regression coefficients imply.

4. Assess the fit of multi-knot spline models

In this step we will consider models in which the relationship between *P* and *Y* is allowed to change at more than one location on *P*. For example, for some fit dimensions a "double-knot spline model" may be more appropriate than the single-knot models described above (e.g., if there is a zone where *P-E* differences are tolerable, or there is lack of agreement among SMEs with regard to *E*). In either case, a double-knot spline model that reflects these possibilities may provide a better fit to the data. Given the level of variation apparent in SMEs' responses to the environment-side measures described in Chapter 13 (ADI and AES), we will evaluate the fit of double-knot spline models that put a confidence interval around estimates for the Army's supply of each fit dimension assessed on the ADI and AES.

5. Select a final model for each fit dimension

One potential criticism of using splines in the manner we propose is that the level of environmental supply of a given attribute is too large a determinant of how the P-E-C relations

[6] In spline regression, a "knot" refers to any place on the regression line where a change in the relationship between the predictor and criterion is modeled.

are modeled. To the extent that one's estimate of E is inaccurate (assuming E is important), one may be losing validity for predicting the criterion by fitting spline models in which the knots reflect some function of E. Therefore, as a final step in the modeling process for each fit dimension, we will take an exploratory look at the relation between person and criterion data using either (a) a smoothing function (e.g., LOESS; Cohen, 1999), (b) a simple polynomial model (the inflection point of this model should be close to E), or (c) a variation on the spline models discussed above that lets the "knot" be a parameter to be estimated by the data (see Marsh & Cormier, 2002, pp. 43-48). If these exploratory models reveal a trend in the data that is consistent with theory (e.g., a V-shaped relation between persons' need for autonomy and probability of attrition), yet the knot (or inflection point) in the *P-Y* relation does not appear where we thought it would be (i.e., it doesn't appear at E), then we will refit a model where a new knot is specified.

It is important to note that we will only use the aforementioned exploratory model as a fit dimension's final model if either (a) we have reason to believe that the value obtained for E might be inaccurate (e.g., due to lack of sufficient SMEs)[7], or (b) the non-linearity in the *P-Y* relation appears to be easily accounted for by an alternative explanation. An example of the latter possibility would be if the relation between a person's need for autonomy and attrition turns positive only once desire for autonomy exceeds supply (E) by one scale point, whereas prior to that point (i.e., $E + 1$) the relation between desire for autonomy and attrition is negative. An added benefit of fitting exploratory, data-driven models in this final step is that it will allow us to compare models based on Steps 1 through 4 to a model where the *P-Y* relation is driven by only the *P-Y* data.

6. Fit an overall model that comprises terms from the final models of each fit dimension

The final step in the modeling process will involve entering the terms from the final model for each fit dimension into one overall model of the criterion. In fitting the models for each fit dimension, we may find that one or more of the fit dimensions has no significant relation with the criterion. If this is the case, we will exclude that dimension from the overall model, particularly if sample size is a concern. Nevertheless, simply eliminating fit dimensions whose models are not statistically significant based on traditional significance levels (e.g., $p < .05$) may lead to exclusion of some fit dimensions that have value when considered along with other fit dimensions (Hosmer & Lameshow, 2000). Thus, at this point in the modeling process, we will err on the side of inclusiveness and include terms for any fit dimension whose final model shows even marginal significance (e.g., $p < .20$) in the initial overall model.

Upon fitting this initial model, it is likely that several terms will fail to reach statistical significance. If this is the case, any number of methods could be used to "prune" the equation. We will adopt a variation on an approach suggested by Edwards (1993). The first step in refining the overall model will be to assess the change in model R^2 that occurs when the terms for a given fit dimension (e.g., the variable reflecting a persons' desire for autonomy and the autonomy spline adjustment term) are removed from the model. If removal of the terms for a given fit

[7] This might be the case for environment-side data gathered for the "future Army" described in Chapter 13. Specifically, only six SMEs provided these data, and they made their ratings based on a necessarily limited description of expected future Army conditions.

dimension does not result in a significant decrement in R^2, then the criterion variance accounted for by that fit dimension is accounted for by the other fit dimensions in the model (i.e., it is "empirically redundant;" Edwards, 1993). Although not mentioned by Edwards, if at the end of this step several fit dimensions are removed from the model, we will assess whether adding the terms for each removed fit dimension back into the model (one dimension at a time) results in a significant increment in R^2. This "check" is important because it could be that the reason the fit dimension was dropped from the full model was because it was empirically redundant with other dimensions that were also dropped. If this were the case, it could be that inclusion of terms from a given fit dimension increments the validity of the reduced model.

Once this check is performed and the reduced model is modified accordingly, the next step involves eliminating any higher-order terms (or, in the case of spline models, spline adjustment terms) that fail to reach statistical significance. As with the case of pruning fit dimensions, if many terms are removed at this step, it would be prudent to assess whether returning those terms to the model (one at a time) results in a significant increase in R^2. Once this check is performed and the reduced model is modified accordingly, the resulting model would be the final model for a predicting a given criterion.

Additional Considerations

Although the steps above outline the process we plan to follow, there are a few unique characteristics of the Select21 research that warrant further discussion: (a) identifying criteria to be modeled, (b) evaluating potential interactions between needs and expectations, and (c) deriving scoring weights for combining person and environment data.

Identifying Criteria to be Predicted

For the concurrent validation, we will focus on modeling multiple criteria with the P-E fit data described in Chapter 13. First, we plan to target the Attrition Cognitions and Re-Enlistment Intentions scales from the Army Life Survey (ALS), and the Future Continuance scale from the Future Army Life Survey (FALS). As noted in Chapter 7, although the ultimate criteria of interest for the P-E fit predictors are attrition and re-enlistment behavior, such data will not be available in the concurrent validation sample. As such, we propose targeting constructs that have been found to be the most proximal antecedents of such behaviors in the literature—intentions or withdrawal cognitions (Griffith, Hom, & Gartner, 2000; Strickland, 2004). Second, in addition to targeting the aforementioned scales from the ALS and FALS, we will also target other scales or composites derived from the ALS/FALS that may be of particular interest to the Army (e.g., satisfaction, commitment, core Army values). The determination of what other ALS/FALS criteria to target will be informed by further analyses of structural relationships among the ALS and FALS scales in both the field test and concurrent validation samples. Once models using each of the ALS/FALS criteria that we target are finalized (using the process outlined above), we will examine the criterion-related validity of the predicted values resulting from these models as predictors of other Select21 criteria (e.g., the other ALS and FALS scales, performance ratings, promotion rate).

In addition to the concurrent validation, we will also examine criterion-related validity evidence for the P-E fit data among recruits in the Select21 attrition database (see Chapter 6). Specifically, we will model the relationship between the P-E fit data and four criteria from this database: (a) BCT attrition (attrition through 2 months of service), (b) AIT attrition (attrition between 2 and 6 months of service), (c) IET attrition (attrition through 6 months of service), and (d) unit attrition through 12 months of service. The process used to model these criteria will be identical to the one outlined above, with the exception that instead of using ordinary least squares regression, we will use logistic regression given the dichotomous nature of attrition criterion.

Evaluating Potential Interactions between Needs and Expectations

For recruits in the Select21 attrition database, we will also explore the joint effects of needs and expectations data on attrition (recall, expectations data will not be gathered from Soldiers in the concurrent validation). As noted in Chapter 13, we hypothesize that expectations will interact with needs and environment data to predict attrition. If the interaction between needs and expectations increment the validity of a "main effects only" model in a way consistent with theory, then we will include the interaction in that fit dimension's final model.

Deriving Scoring Weights

We plan to use regression weights from the final models for each criterion to generate final P-E "predictor scores" for each respondent. Given that such regression weights will be optimized based on the sample in which they are generated (i.e., concurrent validation sample or attrition database sample), it would be desirable to cross-validate the prediction models on which they are based prior to operational use. Ideally, we would create a "hold out" sample in which we could cross-validate the prediction models. Unfortunately, sample sizes for the concurrent validation and attrition analysis database will likely prevent us adopting such a strategy. Nevertheless, other options may be available for assessing the generalizability of the prediction model to other samples.

The fact that we have two independent samples in which we are fitting models (i.e., concurrent validation sample and attrition database sample) provides us with an opportunity to apply the weights obtained from one sample to the other sample. This will be helpful for evaluating how well the prediction models generated in one sample generalize to another sample. As an example, we can apply the weights obtained from modeling attrition cognitions as a function of WVI/ADI data in the concurrent validation sample to WVI/ADI data obtained from recruits in the attrition database. We could then examine the composite that results from applying these weights, and assess whether it has validity for predicting various types of attrition among recruits in the attrition database. Although we cannot generate true cross-validated validity coefficients (because we have different criteria across samples), the criteria of interest (e.g., attrition cognitions and attrition) are conceptually related enough that we expect predictions to generalize (i.e., predictions regarding intentions should generalize to actual behavior). In addition to performing the above analyses, we will also provide formula-based estimates for the population cross-validity of the models we examine (using Cattin's [1980] shrinkage formula). Such coefficients provide an estimate of what the average validity of a model would be across an infinite number of cross-validation samples (Schmitt & Chan, 1998).

APPENDIX J

FAKING RESEARCH RESULTS FOR THE WORK VALUES INVENTORY (WVI)

Table J1. Descriptives Statistics and Fake-Honest Differences for WVI Scale Scores (High Supply Reinforcers)

	Honest		Fake Max		Coached		FM-H		C-H		FM-C
Scale	*M*	*SD*	*M*	*SD*	*M*	*SD*	*d*	*r*	*d*	*r*	*d*
Achievement vs. L	0.69	0.25	0.77	0.20	0.80	0.20	0.34	0.23	0.45	0.28	-0.11
Achievement vs. M	0.52	0.23	0.51	0.19	0.45	0.21	-0.05	0.27	-0.31	0.36	0.28
Advancement vs. L	0.76	0.22	0.79	0.20	0.83	0.19	0.13	-0.12	0.32	0.41	-0.19
Advancement vs. M	0.68	0.22	0.57	0.20	0.57	0.23	-0.52	-0.13	-0.52	0.32	0.00
Co-Workers vs. L	0.58	0.29	0.70	0.23	0.73	0.22	0.39	0.08	0.51	0.36	-0.13
Co-Workers vs. M	0.47	0.30	0.42	0.23	0.38	0.24	-0.18	0.23	-0.30	0.30	0.12
Emotional Development vs. L	0.71	0.27	0.83	0.21	0.88	0.18	0.47	0.20	0.66	0.16	-0.21
Emotional Development vs. M	0.59	0.27	0.74	0.23	0.78	0.22	0.54	0.09	0.71	0.20	-0.17
Feedback vs. L	0.49	0.26	0.67	0.26	0.72	0.21	0.66	0.14	0.84	0.12	-0.18
Feedback vs. M	0.37	0.21	0.37	0.21	0.37	0.19	0.02	0.11	0.01	0.07	0.01
Skill Development vs. L	0.75	0.25	0.83	0.21	0.89	0.17	0.33	-0.04	0.57	0.14	-0.25
Skill Development vs. M	0.68	0.24	0.70	0.20	0.72	0.20	0.09	-0.11	0.18	0.33	-0.10
Security[a] vs. L	0.77	0.21	0.73	0.21	0.78	0.23	-0.20	0.18	0.02	0.19	-0.22
Security vs. M	0.67	0.26	0.46	0.21	0.43	0.23	-0.79	0.09	-0.91	0.25	0.13
Social Service vs. L	0.66	0.30	0.83	0.23	0.87	0.17	0.57	-0.03	0.71	0.11	-0.14
Social Service vs. M	0.51	0.28	0.64	0.25	0.62	0.22	0.48	0.04	0.40	0.28	0.08
Team Orientation vs. L	0.53	0.28	0.75	0.23	0.84	0.19	0.77	0.05	1.11	0.08	-0.36
Team Orientation vs. M	0.40	0.26	0.51	0.22	0.53	0.23	0.45	0.27	0.51	0.12	-0.06
High Supplies Composite (vs. L)	0.66	0.17	0.77	0.17	0.82	0.14	0.62	0.02	0.91	0.19	-0.29
High Supplies Composite (vs. M)	0.54	0.11	0.55	0.10	0.54	0.09	0.04	0.12	-0.03	0.25	0.07
High Supplies Composite (vs. L and M)	0.60	0.12	0.66	0.11	0.68	0.09	0.45	0.04	0.61	0.17	-0.17

Note. n = 186-193 (honest), 91-93 (fake max), and 98-100 (coached). The first score for each dimension reflects the proportion of times the given reinforcer was preferred over reinforcers in the low supply category, and the second score reflects the proportion of times the given reinforcer was preferred over reinforcers in the mid supply category. Composite scores reflect averages across scale scores. d = standardized mean difference, which was calculated by subtracting the honest mean from the faking mean and dividing by the standard deviation from the honest condition. For FM-C d's, the pooled standard deviations of fake max and coached ratings were used as divisors. r = zero-order correlation between honest and faking dimension scores. FM-H and C-H d and r statistics are based on within-subjects data.

[a]This reinforcer was dropped for the field-test version of the WVI.

Table J2. Descriptive Statistics and Fake-Honest Differences for WVI Scale Scores (Low Supply Reinforcers)

	Honest		Fake Max		Coached		FM-H		C-H		FM-C
Scale	*M*	*SD*	*M*	*SD*	*M*	*SD*	*d*	*r*	*d*	*r*	*d*
Autonomy vs. H	0.31	0.26	0.23	0.22	0.23	0.24	-0.30	-0.16	-0.29	0.16	-0.01
Autonomy vs. M	0.36	0.27	0.30	0.23	0.26	0.26	-0.22	0.06	-0.39	0.25	0.18
Comfort vs. H	0.39	0.28	0.20	0.24	0.16	0.18	-0.67	0.23	-0.83	0.10	0.17
Comfort vs. M	0.45	0.29	0.27	0.22	0.20	0.18	-0.61	0.18	-0.87	0.18	0.28
Compensation[a] vs. H	0.40	0.31	0.22	0.25	0.17	0.19	-0.58	0.13	-0.76	0.08	0.19
Compensation vs. M	0.42	0.30	0.26	0.27	0.23	0.20	-0.53	0.05	-0.62	0.20	0.09
Creativity vs. H	0.36	0.25	0.22	0.23	0.16	0.20	-0.57	-0.04	-0.81	0.02	0.24
Creativity vs. M	0.37	0.25	0.25	0.19	0.17	0.18	-0.51	0.12	-0.82	0.06	0.33
Flexible Schedule vs. H	0.35	0.29	0.18	0.23	0.14	0.21	-0.59	0.07	-0.74	0.10	0.16
Flexible Schedule vs. M	0.40	0.28	0.24	0.22	0.20	0.22	-0.59	0.20	-0.75	0.04	0.17
Home vs. H	0.36	0.30	0.28	0.27	0.19	0.21	-0.29	0.04	-0.58	0.16	0.30
Home vs. M	0.38	0.29	0.34	0.26	0.24	0.20	-0.16	0.14	-0.51	0.10	0.36
Independence vs. H	0.20	0.24	0.18	0.23	0.13	0.18	-0.06	0.01	-0.28	0.17	0.21
Independence vs. M	0.22	0.24	0.20	0.21	0.15	0.17	-0.09	0.13	-0.28	0.19	0.20
Influence vs. H	0.35	0.23	0.33	0.23	0.26	0.20	-0.11	0.10	-0.41	0.14	0.30
Influence vs. M	0.40	0.26	0.39	0.21	0.33	0.22	-0.03	0.17	-0.29	0.26	0.28
Variety vs. H	0.31	0.25	0.25	0.21	0.21	0.20	-0.24	0.21	-0.39	0.26	0.16
Variety vs. M	0.34	0.25	0.32	0.20	0.25	0.20	-0.07	0.30	-0.35	0.12	0.29
Low Supplies Composite (vs. H)	0.34	0.17	0.23	0.17	0.18	0.14	-0.61	0.01	-0.91	0.17	0.29
Low Supplies Composite (vs. M)	0.37	0.16	0.29	0.14	0.22	0.13	-0.54	0.13	-0.93	0.25	0.40
Low Supplies Composite (vs. H and M)	0.35	0.16	0.26	0.15	0.20	0.13	-0.60	0.05	-0.95	0.22	0.35

Note. n = 188-193 (honest), 91-93 (fake max), and 98-100 (coached). The first score for each dimension reflects the proportion of times the given reinforcer was preferred over reinforcers in the high supply category, and the second score the proportion of times the given reinforcer was preferred over reinforcers in the mid supply category. Composite scores reflect averages across scale scores. d = standardized mean difference, which was calculated by subtracting the honest mean from the faking mean and dividing by the standard deviation from the honest condition. For FM-C d's, the pooled standard deviations of fake max and coached ratings were used as divisors. r = zero-order correlation between honest and faking dimension scores. FM-H and C-H d and r statistics are based on within-subjects data.

[a]This reinforcer was dropped for the field-test version of the WVI.

Table J3. Descriptive Statistics and Fake-Honest Differences for WVI Scale Scores (Mid Supply Reinforcers)

Scale	Honest		Fake Max		Coached		FM-H		C-H		FM-C
	M	*SD*	*M*	*SD*	*M*	*SD*	*d*	*r*	*d*	*r*	*d*
Ability Utilization vs. H	0.55	0.23	0.48	0.23	0.53	0.21	-0.33	0.15	-0.10	0.21	-0.22
Ability Utilization vs. L	0.71	0.24	0.76	0.22	0.84	0.20	0.20	0.23	0.55	0.31	-0.36
Company Policies/Procedures[a] vs. H	0.55	0.25	0.43	0.24	0.45	0.24	-0.47	0.25	-0.41	0.02	-0.07
Company Policies/Procedures vs. L	0.69	0.23	0.76	0.22	0.82	0.18	0.29	0.16	0.53	0.20	-0.25
Esteem vs. H	0.44	0.22	0.42	0.19	0.39	0.20	-0.10	-0.01	-0.24	0.00	0.15
Esteem vs. L	0.64	0.26	0.73	0.21	0.78	0.19	0.37	0.06	0.55	0.20	-0.19
Fixed Role vs. H	0.37	0.26	0.46	0.26	0.55	0.28	0.35	0.38	0.66	0.26	-0.31
Fixed Role vs. L	0.54	0.27	0.71	0.24	0.84	0.20	0.63	0.21	1.09	0.10	-0.48
Leadership Opportunities vs. H	0.46	0.25	0.56	0.20	0.57	0.22	0.40	-0.01	0.44	0.07	-0.05
Leadership Opportunities vs. L	0.61	0.29	0.79	0.23	0.87	0.17	0.61	-0.02	0.91	0.05	-0.31
Leisure Time vs. H	0.39	0.30	0.16	0.22	0.14	0.20	-0.76	0.04	-0.83	0.01	0.08
Leisure Time vs. L	0.57	0.26	0.41	0.22	0.42	0.23	-0.60	0.13	-0.57	0.36	-0.03
Recognition vs. H	0.36	0.24	0.32	0.24	0.32	0.20	-0.20	0.05	-0.18	0.36	-0.02
Recognition vs. L	0.55	0.28	0.60	0.23	0.72	0.22	0.19	0.18	0.62	0.40	-0.46
Social Status vs. H	0.47	0.27	0.56	0.22	0.52	0.23	0.35	0.25	0.20	0.35	0.16
Social Status vs. L	0.65	0.26	0.80	0.21	0.82	0.20	0.57	0.03	0.66	0.04	-0.10
Societal Contribution vs. H	0.49	0.26	0.66	0.18	0.66	0.22	0.66	0.11	0.66	0.26	0.00
Societal Contribution vs. L	0.66	0.27	0.85	0.18	0.84	0.21	0.68	0.24	0.67	0.23	0.01
Mid Supplies Composite (vs. H)	0.46	0.12	0.45	0.10	0.46	0.09	-0.05	0.11	0.00	0.24	-0.06
Mid Supplies Composite (vs. L)	0.63	0.16	0.71	0.14	0.77	0.13	0.53	0.14	0.92	0.28	-0.39
Mid Supplies Composite (vs. H and L)	0.54	0.10	0.58	0.08	0.62	0.08	0.41	0.24	0.74	0.29	-0.35

Note. n = 187-193 (honest), 91-93 (fake max), and 98-100 (coached). The first score for each dimension reflects the proportion of times the given reinforcer was preferred over reinforcers in the high supply category, and the second score reflects the proportion of times the given reinforcer was preferred over reinforcers in the low supply category. Composite scores reflect averages across scale scores. d = standardized mean difference, which was calculated by subtracting the honest mean from the faking mean and dividing by the standard deviation from the honest condition. For FM-C d's, the pooled standard deviations of fake max and coached ratings were used as divisors. r = zero-order correlation between honest and faking dimension scores. FM-H and C-H d and r statistics are based on within-subjects data.

[a]This reinforcer was dropped for the field-test version of the WVI.

Table J4. Correlations between WVI Scale Scores and the Corresponding ABS Scale Scores

	Correlation with ABS		
WVI Scale/Condition	Honest	Fake Max	Coached
High Supply vs. Low			
Achievement vs. L	.11	**.20**	**.42**
Advancement vs. L	.09	**.21**	.02
Security vs. L	**.15**	**.21**	**.21**
Skill Development vs. L	**.13**	.09	**.38**
Emotional Development vs. L	**.16**	.11	.17
Team Orientation vs. L	**.13**	.15	**.23**
Feedback vs. L	.05	**.22**	**.27**
Co-Workers vs. L	**.15**	**.30**	**.22**
Social Service vs. L	.09	.04	.11
Average *r*	.12	.17	.23
Low Supply vs. High			
Creativity vs. H	-.04	**.34**	.06
Independence vs. H	.05	.00	.04
Variety vs. H	-.01	.11	**.25**
Autonomy vs. H	.06	**.24**	**.43**
Flexible Schedule vs. H	.01	-.01	**.37**
Home vs. H	.10	.10	**.27**
Comfort vs. H	-.04	**.29**	.09
Influence vs. H	.06	**.26**	**.24**
Compensation vs. H	**-.13**	.09	.11
Average *r*	.01	.16	.21

Note. $n = 193$ for ABS-WVI Honest correlations, $n = 93$ for ABS-WVI Fake Max correlations, and $n = 100$ for ABS-WVI Coached correlations. The ABS was administered under normal conditions. WVI-ABS correlations show the relationship between corresponding values and expectations (e.g., WVI Achievement with ABS Achievement). Bolded correlations are statistically significant ($p < .05$, one-tailed).

APPENDIX K

PROCEDURE FOR ESTIMATING CORRELATIONS BETWEEN COPRS AND COMPOSITE SCORES AND THE REMAINING CRITERION MEASURES

As discussed in Chapter 3, each Soldier was rated by 0-3 supervisors (*M* = 1.50, *SD* = 0.70) and 0-8 peers (*M* = 2.60, *SD* = 1.20) during the field test. In the concurrent validation, however, we plan to collect COPRS ratings from one supervisor and three peers for each research participant. Therefore, correlations between COPRS and composite scores and the remaining variables were estimated based on ratings from one supervisor and three peers. The procedure used to estimate these correlations involved four main steps. The first step was to estimate the correlations between supervisor ratings and the other criterion variables (including those not part of the performance model) as if only one supervisor rated each Soldier. To accomplish this, we used the Spearman-Brown formula to estimate the interrater reliability of the supervisor ratings based on the average number of supervisors who rated each COPRS dimension. Next, we computed adjustment/attenuation factors by taking the square root of the ratio of interrater reliability for one supervisor to that for the average number of supervisors in the data set. We then multiplied the correlations between the supervisor ratings and other criterion measures by the adjustment factors. We repeated these steps for the peer COPRS ratings.

The second step was to estimate the standard deviations (*SD*s) for ratings of a single supervisor using a formula provided by Schmidt, Le, and Ilies (2003; equation a11, p. 223). We needed these *SD*s to estimate the covariance matrix in the following step. The *SD*s were estimated using the *SD*s for the mean ratings across all supervisors in the data set and the mean number of supervisors who rated COPRS dimension. This formula was also used to calculate the *SD*s of ratings based on a single peer.

In the third step, we computed covariances between COPRS ratings for one supervisor and one peer (derived from the steps described above) and the remaining criterion measures. Finally, correlations were calculated between the estimated COPRS scores based on ratings of one supervisor and three peers and scores on the other criterion measures. This was accomplished using the formula for linear combinations provided by Nunnally and Bernstein (1994; equation 5-8b, p. 173).